NAVAL RESEARCH LOGISTICS QUARTERLY

MARCH 1976
VOL. 23, NO. 1

OFFICE OF NAVAL RESEARCH

NAVSO P-1278

INFORMATION FOR CONTRIBUTORS

The NAVAL RESEARCH LOGISTICS QUARTERLY is devoted to the dissemination of scientific information in logistics and will publish research and expository papers, including those in certain areas of mathematics, statistics, and economics, relevant to the over-all effort to improve the efficiency and effectiveness of logistics operations.

Manuscripts and other items for publication should be sent to The Managing Editor, NAVAL RESEARCH LOGISTICS QUARTERLY, Office of Naval Research, Arlington, Va. 22217. Each manuscript which is considered to be suitable material for the QUARTERLY is sent to one or more referees.

Manuscripts submitted for publication should be typewritten, double-spaced, and the author should retain a copy. Refereeing may be expedited if an extra copy of the manuscript is submitted with the original.

A short abstract (not over 400 words) should accompany each manuscript. This will appear at the head of the published paper in the QUARTERLY.

There is no authorization for compensation to authors for papers which have been accepted for publication. Authors will receive 250 reprints of their published papers.

Readers are invited to submit to the Managing Editor items of general interest in the field of logistics, for possible publication in the NEWS AND MEMORANDA or NOTES sections of the QUARTERLY.

OPTIMAL INVENTORY POLICY WHEN STOCKOUTS ALTER DEMAND

G. J. Caine
Brown University
Providence, Rhode Island

and

R. H. Plaut
Virginia Polytechnic Institute and State University
Blacksburg, Virginia

ABSTRACT

An inventory model in which future demand is affected by stockouts has been considered recently by B. L. Schwartz. Some generalizations of Schwartz's model are presented in this paper and properties of the optimal policies are determined. In the case of deterministic demand, a set-up cost is included and a mixture of backlogged and nonbacklogged orders is allowed during stockout. It is proved that the optimal policy entails either no stockout or continual stockout, depending on the values of three parameters. For stochastic demand, the effect of stockouts on demand density is postulated, the resulting optimal inventory policy is discussed, and an example involving an exponential density function is then analyzed in detail.

INTRODUCTION

Optimal inventory policies may involve stockouts, even when the demand is assumed to be deterministic. The effect of stockout in inventory models is usually taken into account by means of a stockout (penalty) cost. In some cases this may be appropriate, e.g. when the demand during stockout is met by a priority shipment or extra production run. In other cases, however, stockouts may cause loss of goodwill and affect future demand to the firm. B. L. Schwartz [1, 2] formulated a "perturbed demand" model to analyze this latter situation. The initial results of Schwartz are extended in this paper.

The case of uniform demand rate is considered first. Customer response to stockout is assumed to lower future demand, and the steady state (long-term) situation is analyzed. If there is no restriction on order size or interorder time, it is proved for a generalization of Schwartz's model that the optimal policy entails either no stockout or continual stockout. For problems involving fixed order size, fixed initial inventory level, or fixed interorder time, the equations for the optimal policy are derived.

In the case of stochastic demand, a form for the future demand density as a function of stockouts is proposed. Optimal inventory policy, based upon this assumed form, is discussed and an example involving an exponential density function is presented.

DETERMINISTIC DEMAND

In the case of deterministic demand, consider a firm carrying a commodity for which the potential

demand rate is λ_0 (measured in units per time). That is, the firm would experience demand rate λ_0 if none of its customers were to ever confront a stockout situation. Given that this firm institutes an inventory policy resulting in a fraction α of demand occurring during times of stockout, one would expect to find demand perturbed downward from λ_0 to a lesser rate $\lambda = f(\alpha; \lambda_0)$ as a result of lost customer goodwill. The first practical formulation of this "perturbed demand" concept is attributable to B. L. Schwartz [1, 2].

In short, the implicit assumption in all classical inventory formulations is that demand experienced by an individual firm is insensitive to the firm's operating policy. The firm wishes to choose a policy which reflects an optimal balancing of holding, ordering, salvage and penalty costs. Although a cost is generally associated with stockout, demand is considered independent of the frequency of such occurrences. Intuition, however, must lead one to question this assumption. It surely seems more reasonable in many cases that the loss of goodwill which accompanies repeated stockouts will indeed be manifest through lowered customer demand.

Consider the periodic inventory history depicted in Figure 1. Delivery of an order of size $M-(1-b)L$ signals the start of a period. Inventory depletion occurs at the constant (steady state) rate λ per unit time until, after T units of time have elapsed, a total demand of M units has been registered. In each cycle L units of demand are recorded while stockout conditions prevail and it is assumed that bL of this unsatisfied demand may be backlogged until new supplies become available in the next period, where $0 \leqslant b \leqslant 1$. Such a situation might arise, for example, if some customers were unwilling to wait for the next shipment. (The classical backlog and nonbacklog cases are given by $b=1$ and $b=0$, respectively [3].) Therefore, loss of sales per period totals $(1-b)L$. Once M units have been demanded, a new order of size $M-(1-b)L$ is placed. Delivery is assumed to be immediate and all backlogged demand is filled. Present inventory level now equates with inventory holdings at the previous period's commencement and the cycle repeats.

Suppose the firm wishes to maximize average profit per unit time. Considerable simplification results from observing the sufficiency of an analysis of average profit per unit time in a single cycle. That such a study is indeed sufficient follows from the periodicity of the inventory history. The firm's quest reduces then to the single period problem of calculating an optimal set of values for the decision variables M and L.

Note that periodicity of inventory history is not inconsistent with the perturbed demand hypothesis

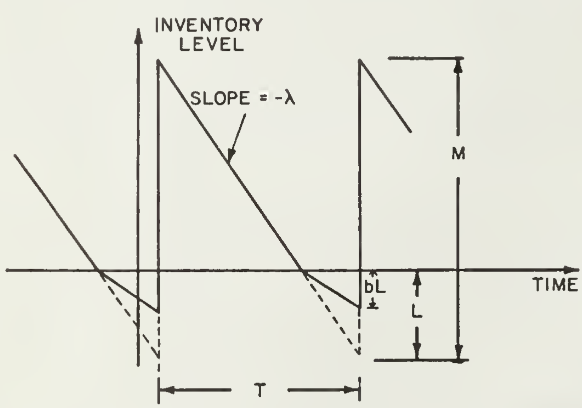

FIGURE 1. Inventory system in steady state with deterministic demand.

that stockouts in one period will influence demand in future periods. For once the firm institutes a permanent inventory policy (by choosing fixed values of M and L), each customer will, in due time, experience a series of satisfactions and disappointments. The system achieves equilibrium when ordering rates have been revised in accordance with the firm's established record of customer service. This study focuses upon the operating characteristics of the stabilized system.

For the moment, consider Schwartz's assumption regarding customer response to stockout. That is, when a customer's demand cannot be immediately satisfied, the customer reacts to this disappointment by purchasing I less units in the future than he would otherwise have purchased. For each unit of stockout the firm loses I units of sales over the infinite future horizon. Schwartz has suggested a method for evaluation of the parameter I [1].

Since the system is operating in equilibrium, the future effect of stockouts experienced in any single cycle must be in balance with the accumulated impact on that cycle of stockouts encountered in previous periods. Therefore, for the purpose of mathematical formulation, it is proper to treat the L stockouts in any period as though they affect the demand rate for that period. It is important to realize that this observation is made solely in the interest of mathematical simplicity. It is not a correct description of the actual dynamics of a system operating under perturbed demand assumptions.

Given a potential demand of λ_0 with no stockouts, $\lambda_0 T$ represents potential demand in a period of length T. Actual demand per cycle, however, equals $\lambda_0 T - LI$ as a result of I units of lost sales accruing to each of the L disappointments per cycle. In view of the fact that total demand per period equals M, it follows that

(1) $$M = \lambda T = \lambda_0 T - LI,$$

or

(2) $$\lambda = \frac{\lambda_0}{1 + (L/M)I} = \frac{\lambda_0}{1 + \alpha I} = f(\alpha : \lambda_0) \cdot$$

This relationship provides the fundamental link between perturbed and potential demand.

We assume that the following cost and revenue factors are operating. A cost per unit time is associated with storing each unit of inventory held. This holding cost is taken to be linear with coefficient H. Ordering cost has two components: in addition to a proportional cost of c per unit ordered, a set-up cost of K is levied (independent of order size) for each order placed. Revenue is proportional to quantity sold with coefficient r. Recall that no immediate penalty arises from stockouts under the assumptions of perturbed demand. Penalties from stockout are reflected in the lowered future demand which derives from the perturbed demand effect.

Since, in every period, the firm sells the same quantity as it orders, it is convenient to formulate the problem in terms of net revenue per unit, defined as $N = r - c$. Therefore profit per cycle, Π, may be written as

(3) $$\Pi = [M - (1-b)L]N - \frac{(M-L)^2 H}{2\lambda} - K.$$

Defining P to be net profit per unit time, it now follows that

$$P=\frac{\Pi}{T}=\frac{\lambda\Pi}{M}=\lambda N\left[1-\frac{(1-b)L}{M}\right]-\frac{(M-L)^2H}{2M}-\frac{K\lambda}{M}. \tag{4}$$

It is always understood that $N>0$, for otherwise the problem is trivial. But $N>0$ implies that $N\lambda_0>0$ so that in finding values of M and L which maximize $P/N\lambda_0$, one has found those values of M and L which maximize P. Therefore it is sufficient to consider the problem of maximizing $P/N\lambda_0$ with respect to M and L.

Define

$$m=\frac{MH}{\lambda_0 N},\qquad l=\frac{LH}{\lambda_0 N},\qquad k=\frac{KH}{\lambda_0 N^2},\qquad p=\frac{P}{\lambda_0 N}. \tag{5}$$

In terms of these nondimensional quantities, (4) becomes

$$p=\frac{[m-(1-b)l-k]}{(m+lI)}-\frac{(m-l)^2}{2m}. \tag{6}$$

The decision variables are l and m, and the relevant parameters have been reduced to k, b, and I.

It is clear that M and L have been defined in such a way that $M\geqslant L\geqslant 0$, which immediately translates into $m\geqslant l\geqslant 0$. Thus, m$=0$ implies $l=0$, or, in other words, the firm is experiencing no stockouts in spite of the fact that it is placing no orders. This circumstance can occur only in the degenerate case of $\lambda=0$. Since the assumption is made in all subsequent formulations that $\lambda>0$, it is only necessary to consider maximization over $m>0$.

Note that if λ is a fixed constant, then l/m assumes a fixed ratio γ with $0\leqslant\gamma\leqslant 1$. An effective approach to the optimization problem consists of first maximizing p along rays in the (l, m) plane (i.e., with γ fixed) and subsequently maximizing over the range of admissible values of γ. For $l=\gamma m$, writing p in terms of γ and m yields the function

$$\bar{p}=\frac{[1-(1-b)\gamma-(k/m)]}{(1+\gamma I)}-\frac{(1-\gamma)^2 m}{2}. \tag{7}$$

Differentiation with respect to m gives

$$\frac{\partial\bar{p}}{\partial m}=\frac{k}{(1+\gamma I)m^2}-\frac{1}{2}(1-\gamma)^2 \tag{8}$$

and

$$\frac{\partial^2\bar{p}}{\partial m^2}=\frac{-2k}{(1+\gamma I)m^3}. \tag{9}$$

Hence $\bar{p}$ is concave with respect to m, and for fixed γ the value of m which maximizes $\bar{p}$ is obtained as

$$\hat{m} = \frac{1}{(1-\gamma)} \sqrt{\frac{2k}{1+\gamma I}} \tag{10}$$

by setting $\partial\bar{p}/\partial m = 0$. Plots of $\hat{m}$ versus γ for parameter values $I=3$ and (a) $k=0.005$, (b) $k=0.28125$, (c) $k=2.0$ are shown in Figure 2.

Note that expression (10) for $\hat{m}$ is independent of b. For a situation in which γ is fixed, (10) gives the value of m which maximizes net profit per unit time. The fact that this optimal value of m is not affected by changes in the parameter b lends an attractiveness to the solution, since a firm may be uncertain about the actual value of this parameter.

Substitution of $m=\hat{m}$ from (10) into (7) leads to the following expression $\hat{p}$ which gives the optimal value of $\bar{p}$ for fixed γ:

$$\hat{p} = \frac{1}{(1+\gamma I)} [1-(1-b)\gamma-(1-\gamma)\sqrt{2(1+\gamma I)k}]. \tag{11}$$

The maximum over all rays is now determined by letting γ vary from 0 to 1. The derivative $d\hat{p}/d\gamma$ can be written in the form

$$\frac{d\hat{p}}{d\gamma} = \frac{1}{(1+\gamma I)^2} \left[(2+I+\gamma I) \sqrt{\frac{(1+\gamma I)k}{2}} + b - 1 - I \right]. \tag{12}$$

Its sign behavior depends only upon the quantity in brackets, which is an increasing function of γ; therefore $d\hat{p}/d\gamma$ is either always positive, always negative, or changes sign once from negative to positive as γ ranges from 0 to 1. It follows that the maximum of $\hat{p}$ occurs at one (or both) of the boundaries $\gamma=0$ or $\gamma=1$. In Figure 2, $\hat{p}$ is plotted versus γ for $b=1$, $I=3$, and (a) $k=0.005$, (b) $k=0.28125$, (c) $k=2.0$. In case (a) $\gamma=0$ is optimum, in (b) both $\gamma=0$ and $\gamma=1$ give the same (maximum) value of $\hat{p}$, and in (c) $\gamma=1$ is optimum.

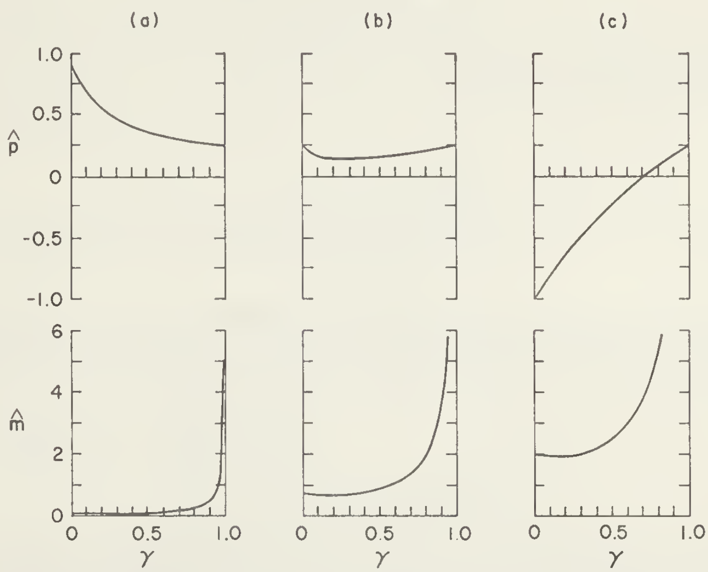

FIGURE 2. Behavior of $\hat{p}$ and $\hat{m}$ for $b=1$, $I=3$, (a) $k=0.005$, (b) $k=0.28125$, (c) $k=2.0$.

For $\gamma=0$, (11) gives $\hat{p}=1-\sqrt{2k}$ and (10) yields $\hat{m}=\sqrt{2k}$, while the case $\gamma=1$ leads to $\hat{p}=b/(1+I)$ and $\hat{m}=+\infty$. Hence the optimal policy is as follows:

$$\text{(13)}\qquad \begin{aligned} &\text{if } 1-\sqrt{2k}\geqslant\frac{b}{(1+I)}, \qquad \text{then } m=\sqrt{2k}, \qquad l=0;\\ &\text{if } 1-\sqrt{2k}<\frac{b}{(1+I)}, \qquad \text{then } m=l=+\infty. \end{aligned}$$

In other words, the firm either should never be out of stock $(l=0)$ or should always be out of stock $(m=l)$. In the first case, note that the optimal order quantity is proportional to the square root of the storage cost. The second case, with $m=l=\infty$, is of course not a practical result. However, other factors or constraints may compensate for this inadequacy, and a few will now be considered.

A firm may wish to restrict the range of γ to $0\leqslant\gamma\leqslant\gamma_0<1$. The arguments above again lead to a boundary solution, and the optimal policy results from a comparison of $\hat{p}$ at $\gamma=0$ and $\hat{p}$ at $\gamma=\gamma_0$. The corresponding value of m is given by (10), and $l=\gamma m$.

It may occur that the nondimensional order quantity m is fixed at some value m_0. Putting $m=m_0$ in (6) and then setting $dp/dl=0$ yields the cubic equation

$$\text{(14)}\qquad (m_0-l)(m_0+lI)^2+kIm_0-(I+1-b)m_0^2=0$$

in the parameter l. Comparison of the net profit p at the real solutions of (14) in the range $0<l<m_0$ and at the boundaries $l=0$ and $l=m_0$ leads to the optimal policy.

Suppose the firm desires that the inventory level at the start of each cycle be a fixed positive value Q_0. Then $m=l+q$ where $q=Q_0H/\lambda_0N$, so that (6) becomes

$$\text{(15)}\qquad p=\frac{bl+q-k}{(1+I)l+q}-\frac{q^2}{2(l+q)},$$

and the condition $dp/dl=0$ implies that

$$\text{(16)}\qquad 2(l+q)^2[(1+I)(k-q)+bq]+q^2[(1+I)l+q]^2=0.$$

This quadratic equation in l can be solved immediately and the optimal policy can be found as in the previous case.

Finally, the interorder time may be prescribed at some value $T=T_0$. It follows from (1) and (5) that $m=t-lI$ where $t=T_0H/N$. Substituting for m in (6) and setting $dp/dl=0$ yields the quadratic equation

$$\text{(17)}\qquad 2(I+1-b)(t-lI)^2-t[t-(1+I)l][(2+I)t-(1+I)lI]=0$$

and the optimal policy can be determined. We note that equations (14), (16), and (17) were previously given (in dimensional form) by Schwartz [2] for the case $K=0$ with $b=0$ and $b=1$.

The results derived above are based on the relationship (2) between perturbed and potential

demand. Other models of customer response are possible, of course; for example, the relationship might be assumed to have the form

$$\lambda=\frac{\lambda_0}{1+[f_1(L)/M]} \tag{18}$$

or

$$\lambda=\lambda_0-[f_2(L)/M] \tag{19}$$

where $f_1(0)=0$ and $f_2(0)=0$. If $b=1$ and if the first derivatives of the functions $f_1(L)$ and $f_2(L)$ are zero at $L=0$, we note that $\partial P/\partial L$ is positive at $L=0$ and hence the optimal policy always involves stockouts. The case $f_2(L)=I_1L^2/2$ has been considered in [4] to model customer response when time of stockout is a factor as well as the amount of stockout.

STOCHASTIC DEMAND

An extension of perturbed demand concepts presented so far is now sought in an effort to analyze the stochastic demand problem. Analogous to the deterministic demand situation, one may view long-term customer density of demand as a downward perturbation (stemming from the firm's operational inventory policy) of potential demand density. For example, allowing interorder time to vary, Schwartz [2, 4] proposes the steady state perturbed demand rate

$$\mu=\frac{\mu_0}{1+\frac{L}{M}I}=\frac{\mu_0}{1+\alpha I} \tag{20}$$

based upon hypothetical demand rate μ_0, disappointment factor $\alpha=L/M$, and fixed parameter I.

Here we will investigate the case of fixed interorder time. It is useful to formulate this problem in terms of an expected disappointment factor. Schwartz [4] considers a long-term expected disappointment factor α defined by

$$\alpha=\frac{\int_y^\infty(\xi-y)\phi_\alpha(\xi)d\xi}{\int_0^\infty\xi\phi_\alpha(\xi)d\xi} \tag{21}$$

where $\phi_\alpha(\xi)$ is the demand density in steady state and y is the stock level at the beginning of each period. Note that $0\leqslant\alpha\leqslant 1$ since demand density functions are zero when their arguments are negative. As an example for $\phi_\alpha(\xi)$, Schwartz gives

$$\phi_\alpha(\xi)=(1+\alpha I)\phi_0[(1+\alpha I)\xi] \tag{22}$$

where I is a positive constant and $\phi_0(\xi)$ is the density of demand experienced each period if no stockouts ever occur.

Equation (21) is a functional relationship in α which may not have an analytical solution and hence may have to be solved numerically for each value of y. In order to avoid this difficulty, we define the alternative expected disappointment factor β,

$$\beta = \frac{\int_y^\infty (\xi - y)\phi_0(\xi)\,d\xi}{\int_0^\infty \xi\phi_0(\xi)\,d\xi} \tag{23}$$

and propose that the long-term demand density $\phi_\beta(\xi)$ be a function of β. Note that β is the ratio of expected stockouts to expected demand, based on the density $\phi_0(\xi)$, and that $0 \leq \beta \leq 1$. Also, we assume that the relationship of $\phi_\beta(\xi)$ to $\phi_0(\xi)$ has the form

$$\phi_\beta(\xi) = g_\beta\phi_0(g_\beta\xi), \qquad g_\beta > 0 \tag{24}$$

where g_β is a function of β. The requirement

$$\int_0^\infty \phi_\beta(\xi)\,d\xi = 1 \tag{25}$$

is automatically satisfied with the form (24).

The expectation associated with ϕ_β satisfies

$$E(\phi_\beta) = \int_0^\infty \xi\phi_\beta(\xi)\,d\xi = \int_0^\infty \xi g_\beta\phi_0(g_\beta\xi)\,d\xi = \frac{1}{g_\beta}E(\phi_0) \tag{26}$$

and, since stockouts are assumed to lower future demand, we must have

$$g_\beta \geq 1. \tag{27}$$

Also, since ϕ_0 is the distribution when there are no stockouts, i.e., when $\beta \to 0^+$, we desire that

$$g_\beta \to 1 \qquad \text{as} \qquad \beta \to 0^+. \tag{28}$$

We assume that g_β is monotonically increasing as β increases from zero.

One might also assume that the long-term demand for the firm's product tends to zero if a stockout situation ($y=0$) always exists. (This is especially true in the nonbacklog situation.) In this case one would require

$$g_\beta \to +\infty \qquad \text{as} \qquad \beta \to 1^-. \tag{29}$$

Some functions satisfying conditions (27)–(29) are

(30) $$g_\beta=\frac{1}{(1-\beta)}\left(1+\beta\sum_{j=0}^{n}c_j\beta^j\right), \qquad c_j\geqslant 0$$

and

(31) $$g_\beta=1-\frac{d}{ln\beta}, \qquad d>0.$$

If the model is to be such that a small amount of stockout (below a threshold $\beta=\beta_0$) has no effect on long-term demand, one might assume a form such as

(32) $$g_\beta=\begin{cases}1 & \text{for } 0\leqslant\beta\leqslant\beta_0\\ g_{\hat\beta} & \text{for } \beta_0<\beta<1\end{cases}$$

where $g_{\hat\beta}$ is given by (30) or (31), say, with β replaced by

(33) $$\hat\beta=\frac{\beta-\beta_0}{1-\beta_0}.$$

The relation

(34) $$g_\beta=1+\beta I, \qquad I>0,$$

analogous to that used in (22), satisfies conditions (27) and (28) but not (29).

As an example, consider the exponential density

(35) $$\phi_0(\xi)=ae^{-a\xi}, \qquad a>0, \qquad \xi\geqslant 0.$$

Then $\phi_\beta(\xi)$is also an exponential density with parameter $g_\beta a$. If g_β is given by (31), one obtains

(36) $$\beta=e^{-ay}, \qquad g_\beta=1+\frac{d}{ay}, \qquad \phi_\beta(\xi)=\left(a+\frac{d}{y}\right)e^{-\left(a+\frac{d}{y}\right)\xi}.$$

A higher value of d implies a larger effect of stockouts on long-term demand.

In order to determine the optimal inventory policy of the firm, consider a period in the steady state. Using the classical theory for now [5], let $\phi(\xi)$ be the density function of demand, x the inventory level before ordering (if $x<0$, then $-x$ denotes the amount of backlogged orders), and y the inventory level after ordering an amount z (so that $y-x=z$, $y\geqslant x$). The purchase or ordering cost is assumed to be $c(y-x)$, the sale price for a unit of stock is r, the holding cost is $h(y-\xi)$ and the stockout (penalty) cost is $p(\xi-y)$†, where the functions c, h, and p are zero for nonpositive arguments. For a given demand ξ, the total loss experienced during the period may then be expressed as

†If the fractional part $(\xi-y)b$ of excess demand may be backlogged, then $p(\xi-y)$ might be assumed to have the form

$$p(\xi-y)=p_1[(\xi-y)b]+p_2[(\xi-y)(1-b)].$$

(37) $$L(y;x)=\begin{cases} c(y-x)+[\min(0,x)]r-[\min(\xi_\nu,y)]r+h(y-\xi_\nu)+p(\xi_\nu-y) & \text{if } y\geqslant 0\\ c(y-x)-(y-x)r+p(\xi_\nu-y) & \text{if } y<0\end{cases}$$

and the expected loss is

(38) $$E[L(y;x)]=$$
$$\begin{cases} c(y-x)+[\min(0,x)]r+\int_0^y[h(y-\xi)-r\xi]\phi(\xi)d\xi+\int_y^\infty[p(\xi-y)-ry]\phi(\xi)d\xi & \text{if } y\geqslant 0\\ c(y-x)-(y-x)r+\int_0^\infty p(\xi-y)\,\phi(\xi)d\xi & \text{if } y<0.\end{cases}$$

For the situation considered in this paper, the effect of stockout is manifest in the density function. In (38), therefore, the stockout cost is deleted and the density $\phi(\xi)$ is replaced by $\phi_\beta(\xi)$ which depends on initial inventory level y. This yields

(39)
$$E[L(y;x)]=\begin{cases} c(y-x)+[\min(0,x)]r+\int_0^y[h(y-\xi)-r\xi]\phi_\beta(\xi)d\xi-ry\int_y^\infty\phi_\beta(\xi)d\xi & \text{if } y\geqslant 0\\ c(y-x)-(y-x)r & \text{if } y<0.\end{cases}$$

If the ordering cost is linear, i.e.,

(40) $$c(z)=zc,$$

then (39) may be written as

(41) $$E[L(y;x)]=-cx+[\min(0,x)]r+G(y)$$

where

(42) $$G(y)=cy+F(y),$$
$$F(y)=\begin{cases}\int_0^y[h(y-\xi)-r\xi]\phi_\beta(\xi)d\xi-ry\int_y^\infty\phi_\beta(\xi)d\xi & \text{if } y\geqslant 0\\ -ry & \text{if } y<0.\end{cases}$$

Now consider the exponential density $\phi_\beta(\xi)$ given in (36). If we assume that the holding cost is linear, i.e.,

(43) $$h(y-\xi)=(y-\xi)h,$$

and define the nondimensional quantities

(44) $$\bar{y}=ay, \qquad \bar{r}=\frac{r}{c}, \qquad \bar{h}=\frac{h}{c}, \qquad \bar{G}=\frac{aG}{c},$$

then (42) yields

(45) $$\bar{G}(\bar{y})=\begin{cases} (1+\bar{h})\bar{y}-\dfrac{(\bar{r}+\bar{h})}{(\bar{y}+d)}\,\bar{y}[1-e^{-(\bar{y}+d)}] & \text{if } \bar{y}\geqslant 0 \\ (1-\bar{r})\bar{y} & \text{if } \bar{y}<0. \end{cases}$$

Note that $\bar{G}(0)=0$, as intuition would suggest. One can show that $\bar{G}''(\bar{y})>0$ for $\bar{y}\neq 0$, and hence $\bar{G}(\bar{y})$ is convex. In Figure 3, plots of $\bar{G}(\bar{y})$ are given for parameter values $\bar{r}=2.0$, $\bar{h}=0.2$, and $d=0.1$, 0.5, 1.0, and 2.0. The minimum values of $\bar{G}(\bar{y})$ occur at $\bar{y}_{\min}=0.567$, 0.405, 0.178, and 0, respectively, and $\bar{y}=\bar{y}_{\min}$ is optimum. The long-term density functions corresponding to the optimal policy for the

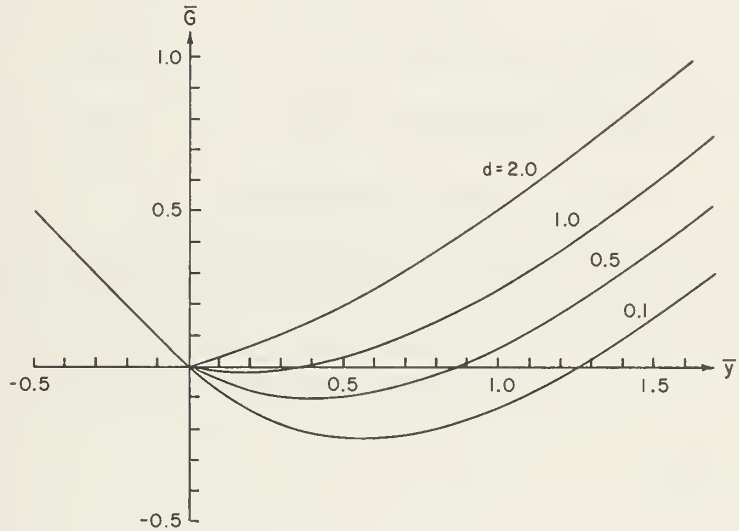

FIGURE 3. Behavior of $\bar{G}$ for $\bar{r}=2.0$, $\bar{h}=0.2$, β formulation.

cases $d=0.1$, 0.5, and 1.0 are shown in Figure 4, along with the hypothetical density function $\phi_0(\xi)$ for $d=0$.

Differentiating (45), one obtains

(46) $$\lim_{\bar{y}\to 0^+}\bar{G}'(\bar{y})=1+\bar{h}-\frac{(\bar{r}+\bar{h})}{d}\,(1-e^{-d})$$

which is positive for sufficiently large values of d. Therefore, if the effect of stockouts on long-term demand is large, then $\bar{y}_{\min}=0$ according to this model and the best policy may be to not carry the item under consideration.

Recall that the disappointment factor β defined in (23) is based on the density $\phi_0(\xi)$. The particular example treated above has been chosen such that one can also obtain analytical expressions with the use of the disappointment factor α defined in (21). If we replace β by α in equations (24) and (31) and

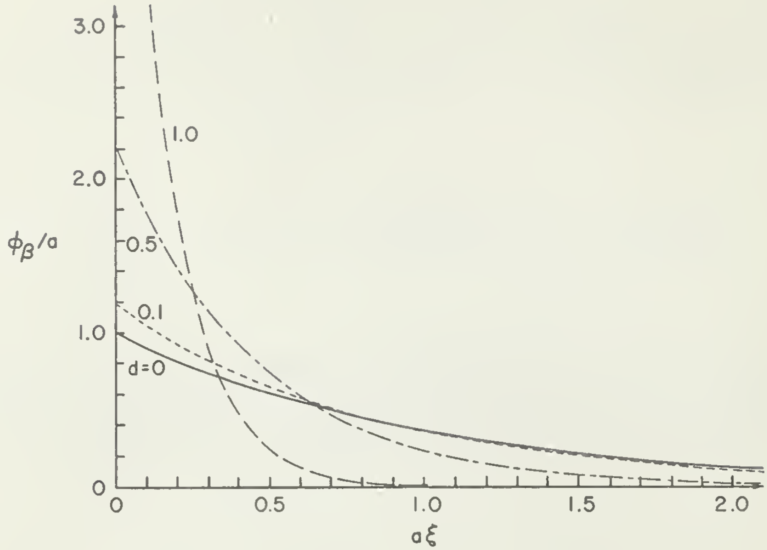

FIGURE 4. Density functions ϕ_β for optimal policies, β formulation.

if $\phi_0(\xi)$ is given by (35), then the relationship (21) becomes

$$\alpha = e^{-[1-(d/ln\alpha)]ay}. \tag{47}$$

Taking logarithms of both sides of (47) leads to a quadratic equation in $ln\alpha$, with the appropriate solution

$$ln\alpha = -1/2(\bar{y} + \sqrt{\bar{y}^2 + 4\bar{y}d}). \tag{48}$$

Using (43), one then obtains

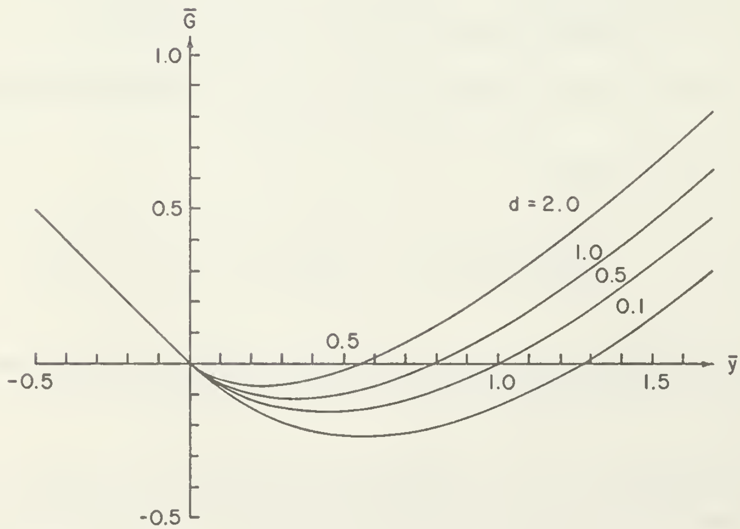

FIGURE 5. BEHAVIOR OF $\bar{G}$ for $\bar{r}=2.0$, $\bar{h}=0.2$, α formulation.

$$\bar{G}(\bar{y}) = \begin{cases} (1+\bar{h})\bar{y} - \dfrac{(\bar{r}+\bar{h})}{[1-(d/ln\alpha)]}\{1-e^{-[1-(d/ln\alpha)]\bar{y}}\} & \text{if } \bar{y} \geqslant 0 \\ (1-\bar{r})\bar{y} & \text{if } \bar{y} < 0 \end{cases} \tag{49}$$

which is plotted in Figure 5 for the same parameter values as in Figure 3. The minimum values of $\bar{G}(\bar{y})$ occur at $\bar{y}_{min} = 0.566$, 0.437, 0.338, and 0.233 for $d = 0.1$, 0.5, 1.0, and 2.0, respectively. All curves have the same slope $(1-\bar{r})$ at $\bar{y} = 0$, so that $\bar{y}_{min} > 0$ unless the unit purchase cost exceeds the unit sale price. The optimal policy is again $\bar{y} = \bar{y}_{min}$.

Comparing Figures 3 and 5, we see that the curves $\bar{G}(\bar{y})$ are similar for the two formulations if the parameter d is small. In other words, the formulation in terms of β and the one based on α lead to similar optimal inventory policies if the effect of stockouts on demand is not too large.\As this effect increases, however, $\bar{y}_{min}$ becomes zero in the β model but remains positive in the α formulation.

REFERENCES

[1] Schwartz, B. L., "A New Approach to Stockout Penalties", Management Science *12*, B–538–B–544 (1966).

[2] Schwartz, B. L., "Optimal Inventory Policies in Perturbed Demand Models", Management Science *16*, B–509–B–518 (1970).

[3] Montgomery, D. C., M. S. Bazaraa, and A. K. Keswani, "Inventory Models with a Mixture of Backorders and Lost Sales", Nav. Res. Log. Quart. *20*, 255–263 (1973).

[4] Schwartz, B. L., "Inventory Models in Which Stockouts Influence Subsequent Demand", Stanford University, School of Business, Research Report No. 12, Stanford, Calif. (1965).

[5] Arrow, K. J., S. Karlin, and H. Scarf, *Studies in the Mathematical Theory of Inventory and Production* (Stanford University Press, Stanford, Calif., 1958).

A MODIFIED BLOCK REPLACEMENT POLICY

Menachem Berg and Benjamin Epstein

TECHNION—Israel Institute of Technology
Haifa, Israel

ABSTRACT

A well known preventive replacement policy is the block replacement policy (BRP). In such a policy the item undergoes a planned replacement at a sequence of equally spaced time points independent of failure history. The main advantage of a BRP is its simplicity, because under this policy it is unnecessary to keep detailed records about times of failures or ages of items. The main drawback of a BRP is that at planned replacement times we may be replacing practically new items. In this paper we study a modified BRP which is free of this drawback. We calculate the expected cost of following a modified BRP for lifetime distributions possessing a special structure and illustrate it for the case of an Erlang distribution. A numerical comparison is made between a modified BRP and a standard BRP for the special case of a two stage Erlang distribution.

1. INTRODUCTION

A preventive replacement policy may be worthwhile in reducing the cost of operating a stochastically failing item. Under such a policy the item is replaced before actual failure (and thus we lose the value of any remaining life) in order to prevent the extra costs associated with a failure.

A well known preventive replacement policy is the block replacement policy (BRP). In such a policy the item undergoes a planned replacement at a sequence of equally spaced time points independent of failure history. If there is more than one item the planned replacement times are common for all of them. This is why the name block replacement is used. The basic BRP model is presented in [1; pp. 95–96]. A working item whose failure is assumed to be immediately detectable, is replaced both at failure and at fixed intervals of time. The replacement is assumed to be instantaneous. The main advantage of this policy lies in its simplicity because no recording of times of failure and ages of items is required. The objective function to be minimized is the minimum average cost per unit of time, for an infinite time horizon.

The main drawback of the BRP is that at planned replacement times we might replace practically new items. The following articles modify the BRP in various ways. In the first model, [1; pp. 96–98], [2], a failed unit is no longer replaced but is instead given a minimal repair. By minimal repair, we mean that the repair, needed to put the failed item back into operation, has no effect on its remaining lifetime. This repair action is mathematically equivalent to replacing the failed item by another working item of the same age. This policy may be appropriate for complex systems because a system undergoing minimal repair can be thought of as a single unit which is aging over time. Bhat [3] also relaxes the requirement in the basic BRP of replacing failed items by new items. In his model failed items are

replaced by items having age t (the block replacement interval), which were taken out unfailed in a former planned block replacement. By reusing these items unused item lifetime is reduced.

In a block replacement model considered by Cox [5], an item which fails "close to" the time of the scheduled block replacement is not replaced and remains idle until block replacement occurs. A penalty, assumed to be a linear function of idle time, is taken into account. Crookes [6] in one of his models (strategy E) follows similar lines. In his model a unit which fails at any time within the interval is not replaced until the next block replacement. Both of these articles contain a mathematical error which has been corrected by Blanning [4]. Woodman [8] suggested the use of dynamic programming to find the optimal policy for the preceding two models.

In this paper we present a different modification of the basic BRP. We call this policy a modified BRP.

2. OUTLINE OF A MODIFIED BRP

The expected cost per unit time per item of a standard BRP taken over the indefinitely long future is

$$C(t) = \frac{c_1 M(t) + c_2}{t}, \tag{1}$$

where t is the length of the interval between scheduled replacements, $M(t)$ is the expected number of failures (unscheduled replacements in $[0, t)$). $M(t)$ is of course the renewal function. c_1 is the cost of making an unscheduled replacement of a failed item. c_2 is the cost per item of a scheduled (block) replacement.

We assume that $c_2 < c_1$, and that the conditional probability of an item failing in the interval $(x, x+\Delta)$ given that it has attained age x is increasing in x. This property is called IFR (Increasing Failure Rate). Mathematically this means that the item failure rate (or hazard rate) $r(x)$ is increasing in x, where

$$r(x) = \frac{f(x)}{\bar{F}(x)}. \tag{2}$$

$f(x)$ and $F(x)$ are respectively the p.d.f. and c.d.f. of the life length of an item and $\bar{F}(x) = 1 - F(x)$.

The principal advantage of a BRP is its simplicity since it renders it unnecessary to keep detailed records about times of failures or ages of items. The principal disadvantage of this policy is its wastefulness because we may replace practically new items at the prescribed replacement points. This led us to consider a modified BRP which is free of this defect.

In the modified BRP we still replace failed items instantaneously after failure, but items possessing age b or less at scheduled block replacement points $t, 2t, 3t, \ldots$ are not replaced by new items but are instead permitted to remain in service; b is a number between 0 and t. Thus at the points $t^+, 2t^+, 3t^+, \ldots$, some of the items will have age zero (0) (following age replacement) and some of the items will have age x, $0 < x \leq b$. We would like to stress that the time points $t, 2t, 3t, \ldots$, are no longer regeneration points as in the ordinary BRP. This makes the mathematical treatment much more complicated and hence new techniques have to be developed. The age distribution in the stationary case is

denoted by the function $\tilde{f}(x)$, which has a discrete point mass for items having age zero and is otherwise continuous for age x in the interval $0 < x \leq b$. Of course

$$\int_{0^+}^{b} \tilde{f}(x)dx + \tilde{f}(0) = 1. \tag{3}$$

For a modified BRP with parameters (b, t), the expected cost per unit time per item taken over the indefinitely long future, is

$$C(b, t) = \frac{c_1 E_x[M_x(t)] + c_2\tilde{f}(0)}{t}, \tag{4}$$

where $M_x(t)$ is the expected number of unscheduled replacements in an interval of length t, if the item is of age x at the start of the interval. Of course $M_0(t) = M(t)$. $E_x[M_x(t)]$ is the expected number of unscheduled replacements in an interval of length t in the stationary case. That is

$$E_x[M_x(t)] = \int_{0^+}^{b} \tilde{f}(x)M_x(t)dx + \tilde{f}(0)M(t). \tag{5}$$

We are interested in finding the values of b and t which minimize $C(b, t)$. To do this we must find $E_x[M_x(t)]$ and $\tilde{f}(0)$.

$M_x(t)$ satisfies the modified renewal equation

$$M_x(t) = F_x(t) + \int_0^t M(t-u)f_x(u)du, \tag{6}$$

where

$$\begin{aligned} F_x(t) &= \frac{F(x+t) - F(x)}{\bar{F}(x)} \\ F_0(t) &\equiv F(t). \end{aligned} \tag{7}$$

$F_x(t)$ is the (conditional) probability that an item having survived to age x will fail to survive for an additional length of time t. $f_x(t) = (d/dt)F_x(t)$ is the (conditional) p.d.f. of the additional life time t of items which have survived to age x. $f_0(t) \equiv f(t)$.

$M(t)$ is obtained by solving the renewal equation

$$M(t) = F(t) + \int_0^t M(t-u)f(u)du. \tag{8}$$

(This is a special case of (6) for $x = 0$).

$\tilde{f}(x)$ and $\tilde{f}(0)$ can be expressed as the unknowns in the following Markovian integral equations

$$\tilde{f}(x) = \int_{0^+}^{b} \tilde{f}(y) p_{yx} dy + \tilde{f}(0) p_{ox}, \qquad 0 < x \leq b$$

(9)

$$\int_{0^+}^{b} \tilde{f}(x) dx + \tilde{f}(0) = 1,$$

where: p_{yx}, $0 < x \leq b$, $0 \leq y \leq b$ is the stationary Markovian transition probability density that an item has age x at the beginning of an interval given that the item had age y at the beginning of the preceding interval.

It can be verified easily that

$$p_{yx} = m_y(t-x)\bar{F}(x), \qquad 0 \leq y \leq b, \; 0 < x \leq b$$

(10)

$$p_{yo} = 1 - \int_{0^+}^{b} p_{yx} dx$$

where

$$m_y(t) = \frac{d}{dt} M_y(t)$$

$$m_0(t) \equiv m(t).$$

Differentiating (6) yields the integral equation

(11) $$m_y(t) = f_y(t) + \int_0^t f_y(u) m(t-u) du.$$

We see from equation (4) that the computation of the cost functions $C(b, t)$ requires knowledge of both $E_x(M_x(t))$ and $\tilde{f}(0)$.

To find them we must know $\tilde{f}(x)$ $(0 < x \leq b)$, which satisfies the integral equation (9). To obtain a general solution for this integral equation is difficult. In the next section we solve $\tilde{f}(x)$ for the case when the item life density, $\tilde{f}(x)$, has a specific structure.

Having sketched the modified BRP it is worthwhile to note that this policy is similar to an optional policy with regular interopportunity replacement intervals, discussed by Woodman in [7]. In the optional policy, t is a given number. In our model, t is a parameter subject to optimization. Woodman, using a dynamic programming technique, presents only the basic functional equations and then solves them for a specific life distribution and specific parameter values. No attempt is made in [7] to derive an analytical solution which gives the cost of a modified BRP as a function of the cost parameters and the item lifetime distribution.

3. CALCULATION OF $C(b, t)$

Assume that $f(t+y)$ can be expressed as

$$f(t+y)=\sum_{i=1}^{k}\tilde{\alpha}_i(t)\tilde{\beta}_i(y). \tag{12}$$

Then it is easy to verify that

$$m_y(t)=\sum_{i=1}^{k}\hat{\alpha}_i(t)\beta_i(y),$$

where

$$\hat{\alpha}_i(t)=\tilde{\alpha}_i(t)+\int_0^t\tilde{\alpha}_i(u)m(t-u)du, \qquad i=1,2,\ldots,k,$$

and

$$\beta_i(y)=\frac{\tilde{\beta}_i(y)}{\bar{F}(y)}.$$

Hence, using (10),

$$p_{yx}=\sum_{i=1}^{k}\alpha_i(x)\beta_i(y), \qquad 0\leqslant y\leqslant b, \qquad 0<x\leqslant b \tag{13}$$

where

$$\alpha_i(x)=\hat{\alpha}_i(t-x)\bar{F}(x).$$

Inserting (13) into (9) yields

$$\tilde{f}(x)=\sum_{i=1}^{k}\alpha_i(x)\left[\int_{0^+}^{b}\tilde{f}(y)\beta_i(y)dy+\tilde{f}(0)\beta_i(0)\right], \qquad 0<x\leqslant b. \tag{14}$$

Hence

$$\tilde{f}(x)=\sum_{i=1}^{k}a_i\alpha_i(x). \tag{15}$$

To find $a_i (i=1,\ldots,k)$ we insert (15) into (14)

$$\sum_{i=1}^{k}a_i\alpha_i(x)=\sum_{i=1}^{k}\alpha_i(x)\left[\int_{0^+}^{b}\beta_i(y)\sum_{j=1}^{k}a_j\alpha_j(y)dy+\tilde{f}(0)\beta_i(0)\right]. \tag{16}$$

Equating the coefficients of $\alpha_i(x)\,(i=1,\ldots,k)$ on both sides of equation (16) yields the set of equations

$$a_i=\sum_{j=1}^{k}a_j\int_{0^+}^{b}\alpha_j(y)\beta_i(y)dy+\tilde{f}(0)\beta_i(0), \qquad i=1,\ldots,k. \tag{17}$$

Recalling that $\tilde{f}(0)=1-\int_{0+}^{b}\tilde{f}(x)dx$ we obtain

$$\tilde{f}(0)=1-\sum_{j=1}^{k}a_j\int_{0+}^{b}\alpha_j(x)dx. \tag{18}$$

Inserting (18) into (17) yields the set of equations for a_i, $(i=1,\ldots,k)$

$$a_i=\beta_i(0)+\sum_{j=1}^{k}a_j\left[\int_{0+}^{b}\alpha_j(y)(\beta_i(y)-\beta_i(0))dy\right],\qquad i=1,\ldots,k. \tag{19}$$

For example for $k=1$ we get

$$a_1=\frac{\beta_1(0)}{1-\int_{0+}^{b}\alpha_1(y)(\beta_1(y)-\beta_1(0))dy}. \tag{20}$$

Let us now define a matrix E

$$E=(e_{ij}),\qquad i=1,\ldots,k,\qquad j=1,\ldots,k \tag{21}$$

where

$$e_{ij}=\int_{0+}^{b}\alpha_j(y)(\beta_i(y)-\beta_i(0))dy,$$

and let $\underline{\beta}(0)$ and $\underline{a}$ be the row vectors

$$\underline{\beta}(0)=(\beta_1(0)),\ldots,\beta_k(0)),\qquad \underline{a}=(a_1,\ldots,a_k). \tag{22}$$

Then (19) can be rewritten as

$$\underline{a}(I-\tilde{E})=\underline{\beta}(0),$$

where $\tilde{E}$ is the transpose of E.
Now if $(I-\tilde{E})$ is a nonsingular matrix then

$$\underline{a}=\underline{\beta}(0)(I-\tilde{E})^{-1}. \tag{23}$$

Hence using (15) we obtain

$$\tilde{f}(x)=\underline{\beta}(0)(I-\tilde{E})^{-1}\bar{\alpha}(x). \tag{24}$$

where $\bar{\alpha}(x)$ is the column vector $\begin{pmatrix}\alpha_1(x)\\ \vdots \\ \alpha_k(x)\end{pmatrix}$.

Hence

$$\tilde{f}(0)=1-\int_{0^+}^{b}\tilde{f}(x)dx=1-\underline{\beta}(0)(I-\tilde{E})^{-1}\int_{0^+}^{b}\overline{\alpha(x)}dx, \tag{25}$$

where $\int_{0^+}^{b}\overline{\alpha(x)}dx$ is the column vector by integrating $\bar{\alpha}(x)$. Inserting (24) and (25) into (5) yields

$$E_x(M_x(t))=\int_{0^+}^{b}M_x(t)\underline{\beta}(0)(I-\tilde{E})^{-1}\bar{\alpha}(x)dx+M(t)\left(1-\underline{\beta}(0)(I-\tilde{E})^{-1}\int_{0^+}^{b}\overline{\alpha(x)}dx\right). \tag{26}$$

Inserting (25) and (26) into (4) yields the required cost function

$$C(b,t)=\frac{c_1\int_{0^+}^{b}M_x(t)\underline{\beta}(0)(I-\tilde{E})^{-1}\bar{\alpha}(x)dx+(c_1M(t)+c_2)\left(1-\underline{\beta}(0)(I-\tilde{E})^{-1}\int_{0^+}^{b}\overline{\alpha(x)}dx\right)}{t}. \tag{27}$$

4. $C(b,t)$ FOR ERLANG DISTRIBUTIONS*

The p.d.f. of an Erlang distribution with m stages is

$$f(t)=\lambda e^{-\lambda t}\frac{(\lambda t)^{m-1}}{(m-1)!},\qquad t\geq 0 \tag{28}$$

and

$$\bar{F}(t)=e^{-\lambda t}\sum_{i=0}^{m-1}\frac{(\lambda t)^i}{i!},\qquad t\geq 0.$$

It can be verified easily that

$$f_y(t)=\frac{f(t+y)}{\bar{F}(y)}=\lambda e^{-\lambda t}\frac{\sum_{i=0}^{m-1}\frac{[\lambda(t+y)]^i}{i!}}{\sum_{i=0}^{m-1}\frac{(\lambda y)^i}{i!}},\qquad t\geq 0 \tag{29}$$

and

$$\bar{F}_y(t)=\frac{\bar{F}(t+y)}{\bar{F}(y)}=e^{-\lambda t}\frac{\sum_{i=0}^{m-1}\frac{[\lambda(t+y)]^i}{i!}}{\sum_{i=0}^{m-1}\frac{(\lambda y)^i}{i!}},\qquad t\geq 0.$$

From [1, p. 57] we have

$$M(t)=\begin{cases}\dfrac{\lambda t}{m}+\dfrac{1}{m}\displaystyle\sum_{j=1}^{m-1}\frac{\theta^j}{1-\theta^j}\left[1-e^{-\lambda t(1-\theta^j)}\right], & m\geq 2\\ \lambda t & m=1\end{cases} \tag{30}$$

*The distribution used as an example in [7] belongs to this family of distributions.

and hence

$$m(t)=\frac{\lambda}{m}\sum_{j=0}^{m-1}\theta^j e^{-\lambda t(1-\theta^j)}, \qquad m\geq 1$$

where $\theta=e^{2\pi i/m}$ is an mth root of unity.

Using the binomial expansion it is easy to verify that $f(t+y)$ in (28) satisfies (12)*. Hence we can use (25) and (26) to compute $\tilde{f}(0)$ and $E_x[M_x(t)]$, respectively.

We now carry out the detailed calculations for the two stage Erlang distribution with $\lambda=1$ (clearly there is no loss of generality since this involves only a scale change).

In this case $m=2$. Hence (28), (29) and (30) become

$$f(t)=te^{-t}, \qquad t\geq 0 \tag{28'}$$

$$\bar{F}(t)=e^{-t}(1+t), \qquad t\geq 0.$$

$$f_y(t)=e^{-t}\frac{t+y}{1+y}, \qquad t\geq 0 \tag{29'}$$

$$\bar{F}_y(t)=e^{-t}\frac{1+t+y}{1+y}, \qquad t\geq 0$$

$$M(t)=1/4(2t-1+e^{-2t}), \tag{30'}$$

$$m(t)=1/2(1-e^{-2t}).$$

Inserting (29′) and (30′) into (6) and (11) we obtain

$$M_y(t)=1/4\left[2t+(1-e^{-2t})\frac{y-1}{y+1}\right], \tag{31}$$

$$m_y(t)=1/2\left(1+e^{-2t}\frac{y-1}{y+1}\right).$$

Using (28′) and (31) to compute p_{yx} (see (10)) we obtain

$$p_{yx}=\frac{e^{-x}(1+x)}{2}\left[1+e^{-2(t-x)}\frac{y-1}{y+1}\right], \qquad 0\leq y\leq b, \qquad 0<x\leq b. \tag{32}$$

Hence

$$p_{yx}=\alpha_1(x)\beta_1(y)+\alpha_2(x)\beta_2(y), \qquad 0\leq y\leq b, \qquad 0<x\leq b$$

where

*In fact the class of distributions based on the Laguerre polynomials (see [9]) which contains the class of Erlang distributions, also satisfies (12).

$$(33) \qquad \alpha_1(x)=\frac{e^{-x}(1+x)}{2}, \qquad \alpha_2(x)=e^{x-2t}(1+x), \qquad \beta_1(y)=1, \qquad \beta_2(y)=\frac{y-1}{2(y+1)}.$$

Hence by (21),

$$(34) \qquad e_{11}=0, \qquad e_{21}=1/2[1-e^{-b}(1+b)], \qquad e_{12}=0, \qquad e_{22}=e^{-2t}[1-e^{b}(1-b)]$$

and

$$(35) \qquad \beta_1(0)=1, \qquad \beta_2(0)=-1/2.$$

Inserting (33), (34), and (35) into (24) yields

$$(36) \qquad \tilde{f}(x)=\frac{e^{-x}(1+x)}{2}-e^{x-2t}(1+x)\frac{e^{-b}(1+b)}{2-2e^{-2t}[1-e^{b}(1-b)]}, \qquad 0<x\leq b.$$

From (33)

$$(37) \qquad \int_{0+}^{b}\alpha_1(x)dx=1-\frac{e^{-b}(2+b)}{2}, \qquad \int_{0+}^{b}\alpha_2(x)dx=be^{-2t+b}.$$

Inserting (34), (35), and (37) into (25) yields

$$(38) \qquad \tilde{f}(0)=e^{-b}\left[1+\frac{b}{2}+\frac{b(1+b)e^{-2t+b}}{2-2e^{-2t}(1-e^{b}(1-b))}\right].$$

Inserting (30′), (31), (36), and (38) into (5) yields

$$(39) \qquad E_x[M_x(t)]=1/4\left[2t-\frac{(1-e^{-2t})e^{-b}(1+b)}{1-e^{-2t}(1-e^{b}(1-b))}\right].$$

Inserting (38) and (39) into (4) yields the required cost function, $C(b,t)$ with

$$(40) \qquad C(b,t)=\frac{c_1}{4t}\left[2t-\frac{(1-e^{-2t})e^{-b}(1+b)}{1-e^{-2t}(1-e^{b}(1-b))}\right]+\frac{c_2}{t}e^{-b}\left[1+\frac{b}{2}+\frac{b(1+b)e^{-2t+b}}{2-2e^{-2t}(1-e^{b}(1-b))}\right].$$

In principle it is possible to find the values of (b^*, t^*) which minimize $C(b, t)$ by computing $\partial C(b,t)/\partial b$ and $\partial C(b,t)/\partial t$, setting them equal to zero and solving the two equations for b^*, t^*. Of course this is not a practical way of finding the optimal values of b^*, t^* and it is simpler to use a computer routine.

It is interesting to compare the minimum cost obtainable with a modified BRP with the minimum cost obtainable with a standard BRP. Let $C(b^*, t^*)$ be the minimum cost for the modified BRP and $C(t_0)$ the minimum cost for a standard BRP. Obviously the block replacement intervals t^* for the modified BRP and t_0 for the standard BRP are in general different.

A numerical analysis of formula (40), where the item life distribution is assumed to be two stage

Erlang, shows that one can save up to 5 percent by using the modified BRP. For $0.25 \leq c_2/c_1 \leq 0.4$, the modified BRP is cheaper than a Failure Replacement Policy (a policy of replacing items only at failure) while the standard BRP is definitely inferior to the FRP (see [1, p. 96]).

It seems reasonable that we could have obtained much higher savings with a modified BRP if we had chosen an Erlang life distribution with many stages.

REFERENCES

[1] Barlow, R. E. and F. Proschan, *Mathematical Theory of Reliability* (John Wiley and Sons, New York, 1965).

[2] Barlow, R. and L. Hunter, "Optimum Preventive Maintenance Policies", Operations Research, *9*, 90–100, (1960).

[3] Bhat, B. R., "Used Item Replacement Policy", Journal of Applied Probability, *6*, 309–318, (1969).

[4] Blanning, R. W., "Replacement Strategies", Operational Research Quarterly, *16*, 253–254, (1965).

[5] Cox, D. R., *Renewal Theory* (John Wiley and Sons, New York, 1962).

[6] Crookes, P. C. I., "Replacement Strategies", Operational Research Quarterly, *14*, 167–184, (1963).

[7] Woodman, R. C., "Replacement Policies for Components that Deteriorate", Operational Research Quarterly, *18*, 267–280, (1967).

[8] Woodman, R. C., "Replacement Strategies", Operational Research Quarterly, *18*, 193–195, (1967).

[9] Zelen, M. and M. C. Dannemiller "The Robustness of Life Testing Procedures Derived from the Exponential Distribution", Technometrics, *3*, 24–49, (1961).

EXPLICIT FORMULAS FOR THE ORDER SIZE AND REORDER POINT IN CERTAIN INVENTORY PROBLEMS

Chandrasekhar Das

University of Saskatchewan
Saskatoon, Canada

ABSTRACT

This study concentrates on distributions of leadtime demand that permit explicit solution to the lot-size, reorder point model. The optimal order size for the general case is first expressed as a function of the economic order quantity and a quantity known as the "residual mean life" in reliability theory. The concept of "no aging" is then utilized to identify a broad class of distributions for which the optimal order size can be determined explicitly, independent of the reorder point.

1. INTRODUCTION

In section 4–2 of their book, Hadley and Whitin [3] discuss a class of static inventory models usually known as the "lot-size, reorder point" models. These models, although approximate in nature, are used frequently by the practitioners because of their simplicity and ease in application. To further facilitate their use, the above authors have developed an iterative technique which is heuristic yet quite efficient for most practical problems under the model. However, use of this technique may in some cases require repeated evaluation of an integral which may not be easy. For this reason we investigate here if in certain of those cases the difficulty in iteration may be avoided and the solution obtained algebraically by using an explicit formula.

2. THE RESULT

For the sake of brevity, we retain the assumptions and notations of Hadley-Whitin to the extent possible in our discussion. This enables us to quote the following results directly from reference [3]:

$$K = \text{expected total system cost per unit time}$$

$$= \frac{\lambda A}{Q} + \left[\frac{Q}{2} + r - \mu\right] IC + \frac{\pi\lambda}{Q}\bar{\eta}(r)$$

where

$$\bar{\eta}(r) = \int_r^\infty x h(x)\,dx - rH(r).$$

Further, the determining equations for the optimal Q and r (that minimize K) are

(1) $$Q=\left\{\frac{2\lambda A}{IC}+\frac{2\pi\lambda\bar{\eta}(r)}{IC}\right\}^{1/2},$$

(2) $$H(r)=\frac{QIC}{\pi\lambda}.$$

Let us now introduce the notation

$$\beta(r)=\bar{\eta}(r)/H(r)$$

so that equation (1) can be rewritten as

$$Q=\left[\frac{2\lambda A}{IC}+\frac{2\pi\lambda\beta(r)H(r)}{IC}\right]^{1/2}.$$

Substituting the right hand side of equation (2) for H(r) in the above expression we then get

(3) $$Q=\{Q_w^2+2\beta(r)Q\}^{1/2}$$

where

$$Q_w=\left(\frac{2\lambda A}{IC}\right)^{1/2}=\text{Wilson's economic lot size.}$$

Hence, squaring both sides of equation (3) and transposing terms

(4) $$Q^2-2\beta(r)Q-Q_w^2=0.$$

Solving the above for $Q>0$, and denoting the optimal values by Q^*, r^* it then follows

(5) $$Q^*=\beta(r^*)+\{\beta^2(r^*)+Q_w^2\}^{1/2}.$$

Under the present model the optimal order size can thus be expressed as an explicit function of $\beta(r^*)$ and Q_w. However, the factor $\beta(r^*)$ which is known as the "mean residual life" function may not be explicit. Therefore, a closed form expression for Q^* can be obtained only if β can be written explicitly in a convenient form. Of course, if the distribution of leadtime demand is such that $\beta(r)$ is independent of r, then Q^* possesses a remarkably simple expression.

We now proceed to illustrate determination of Q^* whether $\beta(r)$ is constant or not.

3. EXAMPLES

Let us first consider the case of uniform distribution of leadtime demand. This distribution is most often used to demonstrate the Hadley-Whittin method; an explicit solution for this case can, however, be obtained as shown below.

Example 1

Let

$$h(x) = (b-a)^{-1}, \qquad a \leq x \leq b.$$

Then it is easily verified that

$$H(r) = (b-r)/(b-a),$$

$$\bar{\eta}(r) = (b-r)^2/\{2(b-a)\},$$

$$\beta(r) = (b-r)/2.$$

From equation (2) we, therefore, get

$$\frac{b-r^*}{b-a} = H(r^*) = \frac{Q^*IC}{\pi\lambda}$$

so that

(6) $$b - r^* = \alpha Q^*$$

where

$$\alpha = \frac{(b-a)IC}{\pi\lambda}.$$

Hence

$$2\beta(r^*)Q^* = (b-r^*)Q^* = \alpha Q^{*2}.$$

Substituting the above in equation (4) and simplifying, we then get

$$Q^* = (1-\alpha)^{-1/2}Q_w.$$

Finally, from equation (6) we derive

$$r^* = b - \alpha Q^* = b - \alpha(1-\alpha)^{-1/2}Q_w.$$

Next we consider a case where $\beta(r)$ is constant.

Example 2

Let

$$h(x) = \theta_2 \cdot \exp(-\theta_1 x), \qquad a \leq x \leq \infty,$$

where θ_1, θ_2 are positive parameters (that may depend on a) such that $h(x)$ is a proper probability distribution. By integration it is easy to verify that

$$H(r) = (\theta_2/\theta_1) \cdot \exp(-r\theta_1),$$

$$\bar{\eta}(r) = (\theta_2/\theta_1^2) \cdot \exp(-r\theta_1),$$

$$\beta(r) = 1/\theta_1.$$

Therefore, from equations (5) and (2)

$$Q^* = (1/\theta_1) + \{(1/\theta_1)^2 + Q_w^2\}^{1/2},$$

$$r^* = (1/\theta_1) \cdot \{\ln(\pi\lambda\theta_2) - \ln(Q^* IC\theta_1)\}.$$

Note that the above formula for Q^* is independent of r^*, although that of r^* involves Q^*. What is more noteworthy is that Q^* does not depend on the penalty cost parameter π. These properties, as proved in Das [2], are special to distributions with constant β. The rest of the paper is, therefore, devoted to a detailed study of this class of distributions.

4. DISTRIBUTIONS WITH CONSTANT β

As mentioned and illustrated above, distributions with constant β enjoy some special properties that give rise to simplification in inventory decision-making. These distributions are also of independent interest in reliability theory since constant β is equivalent to the property of "no aging"—a concept that is valuable in reliability studies. Regarding their characterization, it is widely known that among all continuous distributions defined on the positive half of the real line only the negative exponential distribution exhibits this special property. Example 2, however, shows that the truncated exponential distribution which includes the regular negative exponential as a special case also possesses the same property. For inventory applications of this result we must, of course, verify that the truncated exponential is a realistic distribution of leadtime demand. This is particularly important here because leadtime demand being in general a random sum of random variables has a compound distribution to which the truncated exponential might not be a good fit. However, this doubt is removed, at least in part, by the following example where we allow both the demand per unit time and leadtime to follow a fairly realistic distribution each but the leadtime demand distribution emerges as the truncated exponential.

Example 3

Let us assume that the demand per unit time is distributed normally with mean μ and variance σ^2, and that leadtime is random having a negative exponential distribution with mean γ^{-1}. The conditional distribution of leadtime demand given that leadtime is m, is then normal with mean $m\mu$ and variance $m\sigma^2$:

$$h(x|m) = (2m\pi\sigma^2)^{-1/2} \exp\{-(x - m\mu)^2/2m\sigma^2\}.$$

The marginal distribution of x is, therefore, given by

$$h(x) = \delta \int_0^\infty h(x|m) \exp(-\gamma m) dm, \tag{7}$$

where δ represents the undetermined normalizing factor such that $h(x)$ is a proper probability distribution. The specific numerical value of δ shall depend on the range of x which we prefer to specify later in the paper. Our immediate task is then to evaluate the integral appearing on the right hand side of equation (7). After some simplification it is seen that this integral reduces to

$$(2\pi\sigma^2)^{-1/2} \exp(\mu x/\sigma^2) \cdot I(x) \tag{8}$$

where

$$I(x) = \int_0^\infty m^{-1/2} \exp\{-(x^2 + m^2\rho^2)/2m\sigma^2\} dm,$$

$$\rho^2 = \mu^2 + 2\gamma\sigma^2.$$

To evaluate $I(x)$, let us now make the following transformation

$$y^2 = (\rho^2/2\sigma^2)m.$$

$I(x)$ can then be written as

$$I(x) = 2^{3/2}(\sigma/\rho)J(a) \tag{9}$$

where

$$J(a) = \int_0^\infty \exp[-\{y^2 + (a/y)^2\}]dy,$$

$$a = (\rho/2\sigma^2)x.$$

From the table of integrals (see [4], p. 305, integral no. 427), we next find

$$J(a) = (1/2)\pi^{1/2} \exp(-2a).$$

Hence substituting the above in equation (9) we get

$$I(x) = (2\pi)^{1/2}(\sigma/\rho)\, exp\, (-\rho x/\sigma^2).$$

Therefore, utilizing equation (8) we finally arrive at

$$h(x) = (\delta/\rho) \cdot \exp\{-(\rho - \mu)x/\sigma^2\}$$

which is nothing but the truncated exponential distribution of example 2 with $\theta_2 = \delta/\rho$ and $\theta_1 = (\rho - \mu)/\sigma^2$. In particular, if we assume $0 \leqslant x \leqslant \infty$, then $\delta = \rho\theta_1$ so that $h(x)$ becomes the regular negative exponential distribution.

Similarly, it can be shown that if the demand per unit time itself is negative exponential and the leadtime is geometric, then the leadtime demand is negative exponential. In this context we may also mention Carlson's [1] result: if the leadtime distribution is geometric, then for any arbitrary distribution of demand per unit time with finite cumulants the leadtime demand is asymptotically exponential as the mean leadtime increases.

As regards discrete distributions of leadtime demand, we may point out that the geometric distribution being the discrete analogue of the exponential possesses the "no aging" property. Therefore, pairs of distributions such that their compound is geometric will generate leadtime demand distributions of constant β for the discrete case. An excellent example of this occurs when demand per unit time is negative Binomial and the leadtime is geometric (see Magistad [5]).

5. CONCLUDING REMARKS

The class of leadtime demand distributions permitting explicit formulas for the decision variables under the lot-size, reorder point model is, therefore, not so narrow as might appear to be. Further, if approximate solutions are acceptable, then this class can be augmented by admitting distributions that can be approximated by those with constant β. For instance, Presutti and Trepp [6] show that the standard normal distribution can be approximated by a function of the form $\theta_2 \exp\{-\theta_1|x|\}$, where $\theta_1 = 2^{1/2}$. Hence, if the leadtime demand distribution is normal with arbitrary mean and standard deviation σ, then

$$\beta(r) \simeq \sigma/\theta_1 = (\sigma^2/2)^{1/2}$$

so that

$$Q^* \simeq (\sigma^2/2)^{1/2} + \{(\sigma^2/2) + Q_w^2\}^{1/2}.$$

Besides providing such explicit formulas the function β also facilitates various sensitivity studies of the model discussed here. An application of this nature can be found in Das [2].

REFERENCES

[1] Carlson, P., "On the Distribution of Lead Time Demand," J. of Ind. Eng., *15*, Mar.–Apr., 87–94 (1964).

[2] Das, C., "Sensitivity of the (Q, R) Inventory Model to Penalty Cost Parameter," Naval Res. Log. Quart. *22*, 175–179 (1975).

[3] Hadley, G. and T. M. Whitin, *Analysis of Inventory Systems* (Prentice-Hall Inc., Englewood Cliffs, N.J., 1963).

[4] Hodgman, C. D. et al., *CRC Mathematical Tables* (Chemical Rubber Publishing Company, Cleveland, Ohio, Eleventh Edition, 1957).

[5] Magistad, J. G., "Some Discrete Distributions Associated with Life Testing," Proceedings, Seventh National Symposium on Reliability, and Quality Control, IRE, 1–11 (1961).

[6] Presutti, V. J. and R. C. Trepp, "More Ado About Economic Order Quantities (EOQ)," Naval Res. Log. Quart., *15*, 243–251 (1970).

ON THE EQUIVALENCE OF THREE APPROXIMATE CONTINUOUS REVIEW INVENTORY MODELS

Steven Nahmias

University of Pittsburgh
Pittsburgh, Pennsylvania

ABSTRACT

This paper considers the problem of computing reorder points and order quantities for continuous review inventory systems subject to either a service level constraint or a constraint on the average fraction of time out of stock. It is demonstrated that three apparently distinct models are equivalent under these circumstances. Using this equivalence, streamlined algorithms for computed lot sizes and reorder points are developed.

BACKGROUND

Although the literature abounds with numerous descriptions of inventory models, only a very few of these have ever been applied to a real inventory control problem. Most are too time consuming and complex to be of use in large multiproduct systems, or require too many restrictive assumptions to hold.

Since it is very rare that demands are known with certainty, we will restrict our attention to the situation in which demands are random variables with a known probability distribution. When demands are random, there are essentially two distinct classes of inventory models: periodic review and continuous review. In the periodic review case it is assumed that the state of the system is reviewed at fixed points in time (i.e. periodically) while a continuous review inventory model would be applicable if the state of the system is known at all times.*

Although most industrial inventory control systems fall into the periodic review category, the literature suggests that most scientific inventory systems in use utilize a continuous review methodology. There are a number of reasons for this. Equations for the continuous review case are similar to the familiar lot size equation and hence are more easily understood. In addition, when there is a fixed charge (or set-up cost) for placing an order, optimal policies are quite difficult to compute under the periodic review assumption, and truly effective approximations have yet to be developed. A third reason is, perhaps, that when the review period is relatively small and units are demanded on a one at a time basis, a continuous review approach provides an extremely good approximation.

OPTIMAL POLICY

When units are demanded one at a time and the underlying demand distribution is unchanging with time, it is well known (see for example Hadley and Whitin [4]) that the optimal policy is of the following

*Hadley and Whitin [4] use the term transactions reporting as it indicates that transactions are reported as they take place. However since continuous review appears to be the more accepted term, we will use it.

form—whenever the on hand inventory falls to a level r, an order is initiated for an amount Q to arrive τ units of time into the future. This policy is optimal in that it minimizes the expected average cost over infinitely many ordering cycles.

Computing values of (Q, r) optimally appears to be an extremely difficult problem, and those cases for which solutions do exist are quite restrictive and still require prohibitive computation. We consider two simple approximations that have been suggested in the literature, but before presenting them we require the following definitions. Our notation follows that of Hadley and Whitin [4].

DEFINITIONS AND ASSUMPTIONS

$X(\tau)$ = demand during lead-time τ.

$H(t) = P\{X(\tau) > t\}$ = probability that lead-time demand exceeds t.

λ = average or expected number of units demanded each year.

C = dollar value of each individual unit.

I = cost of carrying one dollar of inventory for 1 year.

A = cost of placing an order.

$n(t) = E\{\max(X(\tau) - t, 0)\}$ = expected number of units which go short during a lead time when the reorder point is t.

In the development of the mathematical model the following assumptions are made:

A1. Costs are charged against:

(a) ordering at $CQ + A$ if $Q > 0$, 0 if $Q = 0$,

(b) holding at IC per unit held per year,

(c) shortage at π per unit short.

A2. There is never more than a single order outstanding.*

A3. The reorder point is $r > 0$.

A4. Excess demand is backlogged.

Based on these assumptions, Hadley and Whitin [4] develop an approximation which requires the simultaneous solution of the two equations:

$$Q = \sqrt{\frac{2\lambda}{IC}(A + \pi n(r))} \tag{1}$$

$$H(r) = \frac{QIC}{\pi\lambda} \tag{2}$$

By modifying their derivation slightly, Wagner [9] has obtained the two equations:

$$Q = \sqrt{\frac{2\lambda A}{IC} + \left(\mu + \frac{2\lambda\pi}{IC}\right) \cdot n(r)} \tag{3}$$

*This assumption may be relaxed if the on-hand plus on-order minus backorder inventory is compared to the trigger point r.

(4) $$H(r)=\frac{2QIC}{IC\mu+2\pi\lambda}.$$

The symbol μ is used to represent the mean lead time demand and is equivalent to $\lambda\tau$.

Both of these systems may be solved fairly efficiently by the following algorithm:

Algorithm I

(a) Choose the initial value of $Q=\sqrt{2\lambda A/IC}$.

(b) Find r from (2) or (4).

(c) Find Q from (1) or (3) using the value of r computed in (b). Stop when successive values of either Q or r are equal or are sufficiently close (that is, within a prespecified tolerance).

In using Hadley and Whitin's model an implicit assumption is that the value of the backorder cost, π, is sufficiently large so that the term $QIC/\pi\lambda$ is always less than one, otherwise (2) would not make sense. Since it is possible that $2QIC/(IC\mu+2\pi\lambda)<1$ while $QIC/\pi\lambda>1$ it can happen that for some small values of π equations (3) and (4) may yield a solution while (1) and (2) will not. This is certainly one advantage of Wagner's approach. However, a recent study by Gross and Ince [3] indicate that in general Hadley and Whitin's model is closer to the optimal than Wagner's a greater portion of the time. The authors speculate that perhaps the Hadley-Whitin model has compensating errors which allow it to perform better in many cases.*

An advantage of these simple models is that they can be modified to allow for a variety of generalizations including incremental or all units quantity discounts, orders which must be a multiple of a fixed batch size, problems with space or budgetary constraints, and cases where the lead time, τ, is not known with certainty but is a random variable. Hadley and Whitin discuss the modifications necessary in the computations to deal with situations of this type.

SERVICE LEVELS

A serious problem with actually applying the two models defined above is that it is often difficult if not impossible to assign dollar values to the cost of shortage. As an alternative, one may specify either the service level (probability of a stockout occurring during lead time) or the average time out of stock (probability that any unit demanded can be satisfied with available stock). The latter is also referred to as the fill rate. Although it is evident that there exists a great deal of confusion regarding the distinction between these two criteria, their difference is discussed at length in a number of places. (Brown [1] is one example.) Gross, Harris and Robers [2] present an algorithm for computing (Q, r) values using Wagner's model (equations (3) and (4)) when the service level is specified. The same method can be used for Hadley and Whitin's model. If we can assume that α is the prespecified service level then the method is:

Algorithm II

(a) Let the initial value of $Q=\sqrt{2\lambda A/IC}$.

(b) Find r by solving $H(r)=\alpha$.

(c) Compute π from (2) (or (4)).

*I am indebted to the referee for bringing this reference to my attention.

(d) Find $n(r)$*.

(e) Compute the new value of Q from (1) (or (3)).

Return to (c) or terminate computations when successive values of Q or r are equal or are sufficiently close.

The same approach can be extended to develop an algorithm for computing (Q, r) values when the average fraction of time spent out of stock is specified. Hadley and Whitin describe the following procedure for computing the optimal (Q, r) values subject to the constraint $n(r)/Q=\beta$:

Algorithm III

(a) Choose an initial value of $\pi=\pi_0$.

(b) Using π_0 find the optimal values of (Q, r) solving (1) and (2) (or (3) and (4)). This involves iterating successive values of (Q, r) as described in Algorithm I.

(c) For the given solutions obtained in (b) compute $n(r)/Q$. If $n(r)/Q > \beta$ choose $\pi > \pi_0$ and return to (b). If $n(r)/Q < \beta$ choose $\pi < \pi_0$ and return to (b). Stop only when $n(r)/Q$ is sufficiently close to β.

Anyone who has had any experience with this algorithm can attest to the fact that it is extremely time consuming. Choosing the proper values of π at each stage is largely a hit and miss affair, and to obtain an accurate solution may require as many as 400 computations.

We will demonstrate that both of these algorithms can be streamlined considerably with no loss in accuracy. The value of π may be completely eliminated from equations (1) and (2) as follows: From equation (2) we have that

$$\pi=QIC/\lambda H(r).$$

Substituting this into equation (1) we obtain:

$$Q=\sqrt{\frac{2\lambda}{IC}\left\{A+\frac{QIC}{\lambda H(r)}\cdot n(r)\right\}}$$

which can be seen to be a quadratic equation in Q. The positive root obtained from the quadratic formula is:

$$Q=\frac{n(r)}{H(r)}+\sqrt{\left(\frac{n(r)}{H(r)}\right)^2+\frac{2\lambda A}{IC}} \tag{5}$$

It is surprising to note that had we solved for π in (4) and substituted into (3) we would still obtain equation (5) as the solution.† When the lead time demand distribution is exponential or is approximated

---

* $n(r)=\int_r^\infty xh(x)dx-rH(r)$ where $h(x)=\frac{dH(x)}{dx}$

When lead time demand is normally distributed, $n(r)$ may be obtained as a function of the standardized loss integral. See Hadley and Whitin [4], Wagner [9] or Brown [1].

†That these two apparently different sets of equations yield identical (Q, r) values when a service constraint is imposed is quite surprising. What will be different is the imputed cost of shortage, i.e. the value of π obtained from Algorithms II or III.

by an exponential tail, the variable r drops out of equation (5) so that the order quantity may be determined independently of the reorder point. This special case is discussed by Parker [6], Presutti and Trepp [7] and Schroeder [8]. Brown [1] has also obtained equation (5) by a different derivation. Hence we see that when dealing with service constraints, the models of Hadley and Whitin (equations (1) and (2)), Wagner (equations (3) and (4)) and Brown (equation (5)) are all identically equivalent.

The use of equation (5) completely eliminates the need of an algorithm to determine the optimal (Q, r) values when a service level constraint is imposed. If we let α= probability of stockout during leadtime, then the reorder point r is computed to satisfy the equation

$$H(r) = \alpha.$$

Since specification of r determines $n(r)$, the optimal lot size, Q, may be computed directly from equation (5) (with α substituted for $H(r)$). The optimal (Q, r) values obtained in this manner will be identical to those computed from Algorithm II independently of which model is used. By eliminating the need for an algorithmic solution, the use of equation (5) can reduce computation time by as much as a factor of 10 or 20.

Equation (5) can also be used to significantly reduce computation time for the more common case when β is specified as the average fraction of time spent out of stock. In that case the following algorithm would be utilized:

Algorithm IV

(a) Pick an initial value of $Q = \sqrt{2\lambda A/IC}$.

(b) Compute r from the equation $n(r) = \beta Q$.

(c) Compute Q from (5). Return to (b) or terminate computations when successive values of Q or r are equal or are sufficiently close.

Algorithm IV will yield the identical (Q, r) values as those obtained from Algorithm III, but with a very small fraction of the computations. In fact, depending upon the accuracy desired, Algorithm IV may require as little as 1 percent of the computation time. This is certainly a significant difference especially when considering that such computations might be performed repeatedly for thousands of items on a continuing basis. Note also that once the optimal values of (Q, r) are determined, the imputed cost of a backorder may be computed from either equations (2) or (4).

In addition to the fact that equation (5) significantly streamlines previous algorithms, there are a number of interesting conclusions which can be drawn from the equivalence which we have demonstrated:

(1) Although the formulas presented by Brown [1], Hadley and Whitin [4] and Wagner [9] appear to be different, all three models will yield identical results when a constraint is placed on either the probability of stocking out during lead time or the average fraction of time out of stock.

(2) Equation (5) remains valid for any lead time demand distribution, $H(r)$. It is not necessary to assume an exponential approximation.

(3) Since equation (5) obtained by Brown [1] is equivalent to two approximate models, it itself must be an approximation.

(4) The use of equation (5) significantly reduces computation time over Algorithms II and III when either α (service level) or β (fraction of time out of stock) are specified.

Acknowledgement

The author gratefully acknowledges the comments of Professor Edward Silver on an earlier draft.

REFERENCES

[1] Brown, R. G., *Decision Rules for Inventory Management*, (Holt, Rinehart and Winston, New York, 1967).

[2] Gross, D., C. Harris, and P. Robers, "Bridging the Gap Between Mathematical Inventory Theory and the Construction of a Workable Model," International Journal of Production Research, *10*, 201–214, (1972).

[3] Gross, D. and R. Ince, "A Comparison and Evaluation of Approximate Continuous Review Inventory Models," International Journal of Production Research, *13*, 9–23, (1975).

[4] Hadley, G. and T. Whitin, *Analysis of Inventory Systems*, (Prentice Hall, New Jersey, 1963).

[5] Herron, D., "Service Level vs. Stockout Penalties—A Suggested Synthesis," Production and Inventory Management, *10*, 43–53, (1969).

[6] Parker, L., "Economical Reorder Quantities and Reorder Points with Uncertain Demand," Nav. Res. Log. Quart., *11*, 351–358, (1964).

[7] Presutti, V. and R. Trepp, "More Ado About EOQ," Nav. Res. Log. Quart.," *17*, 243–251, (1970).

[8] Schroeder, R., "Managerial Inventory Formulations With Stockout Objectives and Fiscal Constraints," Nav. Res. Log. Quart., *21*, 375–388, (1974).

[9] Wagner, H. M., *Principles of Operations Research*, (Prentice Hall, New Jersey, 1969).

MULTIPERIOD CAPACITY EXPANSION AND SHIPMENT PLANNING WITH LINEAR COSTS

C. O. Fong

Faculty of Economics and Administration
University of Malaya
Kuala Lumpur, Malaysia

and

V. Srinivasan

Graduate School of Business
Stanford University

ABSTRACT

In this paper we consider a multiperiod deterministic capacity expansion and shipment planning problem for a single product. The product can be manufactured in several producing regions and is required in a number of markets. The demands for each of the markets are nondecreasing over time and must be met exactly during each time period (i.e., no backlogging or inventorying for future periods is permitted). Each region is assumed to have an initial production capacity, which may be increased at a given cost in any period. The demand in a market can be satisfied by production and shipment from any of the regions. The problem is to find a schedule of capacity expansions for the regions and a schedule of shipments from the regions to the markets so as to minimize the discounted capacity expansion and shipment costs. The problem is formulated as a linear programming model, and solved by an efficient algorithm using the operator theory of parametric programming for the transporation problem. Extensions to the infinite horizon case are also provided.

1. INTRODUCTION

The deterministic multiperiod, multiproducing region capacity expansion problem with concave capacity expansion costs was first proposed by Manne [4]. Exact optimal solutions to the problem can be obtained by using dynamic programming formulations [1, 2, 4]. However, owing to the "curse of dimensionality", a problem with more than three producing regions is computationally prohibitive.

In this paper we make the assumption that the capacity expansion costs are linear rather than concave. Linear capacity expansion costs may be realistic in the case of firms which rent or subcontract their production capacities from other firms or when the fixed components in the capacity expansion costs are relatively small. Such a formulation is also realistic in multiperiod personnel assignment problems with different types of jobs as markets and different training programs defining the producing regions. In such contexts, the cost of recruiting and training may be best approximated as linear. Moreover, the optimum solution to the problem with linear costs can be used to provide a lower bound to the objective function of the problem with concave expansion costs. This information is useful in a branch and bound approach to solving the problem under concave expansion costs, especially if the linear prob-

lem can be solved efficiently. Further, studying the structure of the problem with linear costs may provide a basis for the development of more efficient heuristics as aids in solving the problem with concave expansion costs.

In the next section, the linear capacity expansion and shipment problem is formulated and in section 3 we develop a solution procedure to solve it. In section 4 some characterizations of the optimal solution as provided by the algorithm developed in section 3 are derived. In the same section we extend the formulation to the case of infinite planning horizon and show how an optimal solution to that case can be obtained.

2. PROBLEM FORMULATION

We consider below the multiperiod capacity expansion and shipment problem ($P1$) for a single product by defining the index sets:

$I=\{1, 2, \ldots, m\}$	set of producing regions,
$J=\{1, 2, \ldots, n\}$	set of markets,
$K=\{1, 2, \ldots, T\}$	set of time periods where T is the end of the planning horizon and,
$K'=\{2, 3, \ldots, T\}$.	

For $i \epsilon I$, $j \epsilon J$ and $t \epsilon K$, we define

r_j^0 = initial demand in market j, in time period 0.

r_j^t = known increment in market j's demand in time period t. $r_j^t \geqslant 0$. Consequently, $\sum_{\tau=0}^{t} r_j^\tau$ represents the demand in market j at time t.

q_i^0 = initial production capacity in region i. Capacities are measured in the same units as demands.

q_i^t = cumulative amount of capacity added in region i from period 1 to t. The total production capacity in region i at time t is $q_i^0 + q_i^t$.

k_i = unit cost of capacity expansion in region i ($k_i > 0$).

x_{ij}^t = amount shipped from region i to market j during period t.

c_{ij} = unit cost of shipping from region i to market j. This consists of (a) the cost of transporting one unit from region i to market j, (b) the variable cost of producing one unit in region i, (c) maintenance cost for one unit of capacity in region i. ($c_{ij} \geqslant 0$)

s_i^t = excess (or idle) capacity in region i at time t ($\geqslant 0$).

h_i = unit cost of maintaining idle capacity in region i. Obviously $0 \leqslant h_i \leqslant c_{ij}$ for $i \epsilon I, j \epsilon J$.

g_i = terminal (or resale) value of a unit capacity at the end of planning horizon in region i ($g_i \geqslant 0$).

α = discount factor per period $\left(=\frac{1}{1+\mathscr{I}} \text{ where } \mathscr{I} \text{ is the appropriate cost of capital per period}\right)$ $0<\alpha<1$. Thus α^{t-1} is the present value (i.e., value as of the beginning of period 1) of one dollar at the beginning of time period t.

The problem $P1$ can be formulated as below:

$$(1) \quad \text{Min} \sum_{t \epsilon K} \sum_{i \epsilon I} \sum_{j \epsilon J} \alpha^{t-1} c_{ij} x_{ij}^t + \sum_{t \epsilon K} \sum_{i \epsilon I} \alpha^{t-1} h_i s_i^t + \sum_{i \epsilon I} k_i q_i^1 + \sum_{t \epsilon K'} \sum_{i \epsilon I} \alpha^{t-1} k_i (q_i^t - q_i^{t-1}) - \sum_{i \epsilon I} \alpha^T g_i q_i^T$$

(2) $$\sum_{j \in J} x_{ij}^t + s_i^t - q_i^t = q_i^0 \qquad \text{for } i \in I,\ t \in K,$$

(3) $$\sum_{i \in I} x_{ij}^t = \sum_{\tau=0}^{t} r_j^\tau \qquad \text{for } j \in J,\ t \in K,$$

(4) $$q_i^t - q_i^{t-1} \geq 0 \qquad \text{for } i \in I,\ t \in K', \text{ and}$$

(5) $$x_{ij}^t \geq 0,\ s_i^t \geq 0,\ q_i^t \geq 0 \qquad \text{for } i \in I,\ j \in J,\ t \in K$$

where x_{ij}^t, s_i^t, q_i^t are the decision variables.

Objective function (1) determines the minimum total time discounted transportation, idle capacity maintenance and capacity expansion costs. The costs are assumed to be incurred at the beginning of each period. Constraint (2) states that the amount shipped out of region i at time t should not be greater than the capacity available in region i at time t. It is assumed that the amount of capacity addition in region i at time t, $q_i^t - q_i^{t-1}$ (at $t=1$, it is q_i^1), is available for production to meet the demands in the same period t. Constraint (3) states that the amount shipped to market j at time t should be equal to the demand. Constraint (3) implicitly assumes that no backlogging of the product or inventory for the future is permitted. This is quite a realistic assumption since the unit time period involved in these problems is usually a year. Since $h_i \leq c_{ij}$ for $i \in I$, it is obvious that stating constraint (3) as equality (rather than as $\geq$) involves no loss of generality. Constraint (4) follows from the fact that q_i^t expresses the cumulative amount of capacity expansion until period t.

As pointed out in [3], there is no established precedent regarding the method to be employed in computing the terminal value of capacity. The life of a production capacity is usually much longer than a normal planning horizon. Manne [4] in many of his studies assumed an infinite life for capacity. Under this assumption of no deterioration in the manufacturing production capability over time, we posit that the capacity can be sold off approximately for its cost value (i.e., $k_i q_i^T$) at the end of the time horizon. Therefore it is not unrealistic to assume that the present value of the terminal value of the capacities is $\sum_{i \in I} \alpha_i^T k_i q_i^T$. (For the infinite horizon case that will be considered later, this assumption is not needed.) Objective function (1) can now be modified to

(6) $$\text{Min} \sum_{t \in K} \sum_{i \in I} \sum_{j \in J} \alpha^{t-1} c_{ij} x_{ij}^t + \sum_{t \in K} \sum_{i \in I} \alpha^{t-1} h_i s_i^t + \sum_{i \in I} k_i q_i^1 + \sum_{t \in K'} \sum_{i \in I} \alpha^{t-1} k_i (q_i^t - q_i^{t-1}) - \sum_{i \in I} \alpha^T k_i q_i^T$$

Regrouping the cost coefficient of each q_i^t in (6) we simplify (6) to (7).

(7) $$\text{Min} \sum_{t \in K} \sum_{i \in I} \sum_{j \in J} \alpha^{t-1} c_{ij} x_{ij}^t + \sum_{t \in K} \sum_{i \in I} \alpha^{t-1} h_i s_i^t + \sum_{t \in K} \sum_{i \in I} \alpha^{t-1} k_i' q_i^t$$

where $k_i' = (1-\alpha) k_i$.

Under this formulation, the company is, in effect, renting at the rate $\mathscr{I}$ capacities worth $k_i q_i^1, k_i q_i^2, \ldots, k_i q_i^T$ in periods $1, 2, \ldots, T$ so that the present value of the rent expenses would be

$$\sum_{i\epsilon I}\left\{\frac{[\mathscr{I}(k_iq_i^1)]}{(1+\mathscr{I})}+\frac{[\mathscr{I}(k_iq_i^2)]}{(1+\mathscr{I})^2}+\ .\ .\ .\ +\frac{[\mathscr{I}(k_iq_i^T)]}{(1+\mathscr{I})^T}\right\}$$

since $\mathscr{I}/(1+\mathscr{I})=(1-\alpha)$ the above expression is the same as the last term in (7). Consequently, (7) minimizes the intuitively appealing objective defined as the sum of total time discounted transportation costs, idle capacity maintenance costs, and the rental costs on the production capacities.

The linear cost capacity expansion and shipment problem can now be formulated as problem $P2$, which is minimizing (7) over constraint set (2)-(5).

$P2$ can be solved using the simplex method for linear programs. However, for any reasonably large size problem, the number of constraints and variables will be quite large. For example, a 10-period, 10-region, 200-market problem will have 2,190 constraints (excluding the non-negativity constraints) and 20,200 variables making it computationally expensive.

3. SOLUTION PROCEDURE

3.1. Development

To develop an efficient solution procedure for $P2$, we consider the problem $P3$ of minimizing (7) over constraint set (2), (3), and (5) (i.e., after dropping the coupling constraints (4)). Problem $P3$ consists of a series of subproblems $P3^t$ for $t\epsilon K$. Our solution procedure is such that the optimal solutions for subproblems so obtained also satisfy the coupling constraints (4) so that they are also optimal to $P2$. Each $P3^t$ is of the form:

(8) $$\text{Min} \sum_{i\epsilon I}\sum_{j\epsilon J}\alpha^{t-1}c_{ij}x_{ij}^t+\sum_{i\epsilon I}\alpha^{t-1}h_is_i^t+\sum_{i\epsilon I}\alpha^{t-1}k_i^tq_i^t$$

(9) $$\sum_{j\epsilon J}x_{ij}^t+s_i^t-q_i=q_i^0 \qquad \text{for } i\epsilon I,$$

(10) $$\sum_{i\epsilon I}x_{ij}^t=\sum_{\tau=0}^{t}r_j^\tau \qquad \text{for } j\epsilon J, \text{ and}$$

(11) $$x_{ij}^t\geqslant 0,\ s_i^t\geqslant 0,\ q_i^t\geqslant 0 \qquad \text{for } i\epsilon I,\ j\epsilon J.$$

To bring each $P3^t$ to "standard form", we convert each to a capacitated transportation problem [5] with $m+1$ regions and $n+2$ markets.

We define a dummy $(n+1)$ market with

$$r_{n+1}^0=\max\left\{0,\sum_{i\epsilon I}q_i^0-\sum_{j\epsilon J}r_j^0\right\} \text{ and } \sum_{\tau=0}^{t}r_{n+1}^\tau=\sum_{i\epsilon I}q_i^0 \text{ for } t\epsilon K,$$

to absorb the excess capacities.* It is shown below that defining $\sum_{\tau=0}^{t}r_{n+1}^\tau=\sum_{i\epsilon I}q_i^0$ for any $t\epsilon K$ is sufficient to absorb the excess capacities in an optimal solution to $P3^t$.

* In other words, $r_{n+1}^1=\sum_{i\epsilon I}q_i^0-r_{n+1}$ and $r_{n+1}^\tau=0$ for $1<\tau\leqslant T$.

LEMMA 1: In an optimal solution to $P3^t$, $t\epsilon K, s_i^t \leqslant q_i^0$ for $i\epsilon I$.

PROOF: Assume the contrary, i.e., $\exists$ a k such that $s_k^t > q_k^0 \geqslant 0$. But $\sum_{j\epsilon J} x_{kj}^t + s_k^t - q_k^t = q_k^0$ and hence we have $q_k^t > 0$. Consequently, the objective function to $P3^t$ can be reduced by decreasing q_k^t and s_k^t, contradicting the optimality assumption.

COROLLARY 1: In an optimal solution to $P3^t$, $\sum_{i\epsilon I} s_i^t \leqslant \sum_{i\epsilon I} q_i^0$.

The excess "demand" of the $(n+1)$ markets can be satisfied by defining a dummy $(m+1)$ region with $q_{m+1}^0 = \max\left\{0, \sum_{j\epsilon J} r_j^0 - \sum_{i\epsilon I} q_i^0\right\}$, $c_{m+1,\ j} = M$(a large positive number) for $j\epsilon J$, $h_{m+1} = c_{m+1,\ n+1} = 0$ and $k_{m+1} = 0$, without increasing the value of the objective function. Consequently, there is a one-to-one correspondence between an optimal solution to $P3^t$ and an optimal solution to $P3^t$ with the addition of the $(m+1)$ region and $(n+1)$ market.

Hence, we can incorporate the following constraint to $P3^t$,

(9a)
$$\sum_{j\epsilon J} x_{m+1,j}^t + s_{m+1}^t - q_{m+1}^t = q_{m+1}^0.$$

Eq. (10) is modified to:

(10a)
$$\sum_{i\epsilon I} x_{ij}^t + x_{m+1,j}^t = \sum_{\tau=0}^{t} r_j^\tau \qquad \text{for } j\epsilon J \text{ and}$$

(10b)
$$\sum_{i\epsilon I} s_i^t + s_{m+1}^t = \sum_{\tau=0}^{t} r_{n+1}^\tau.$$

Define $I' = I \cup \{m+1\}$, $J' = J \cup \{n+1\}$ and $J'' = J' \cup \{n+2\}$.

Following the method of [8], we add N (a very large positive integer) to both sides of (9) and (9a) to obtain,

(9b)
$$\sum_{j\epsilon J} x_{ij}^t + s_i^t + (N - q_i^t) = q_i^0 + N \qquad \text{for } i\epsilon I'.$$

Summing (9b) over $I\epsilon I'$ and subtracting from it (10b) and the sum of (10a) over $j\epsilon J$ we have:

$$\sum_{i\epsilon I'} (N - q_i^t) = (m+1)N + \sum_{i\epsilon I'} q^0 - \sum_{j\epsilon J'} \sum_{\tau=0}^{t} r_j^\tau.$$

It can be shown in a similar manner as in [8] that each $P3^t$ is equivalent to $P4^t$, which is

(12)
$$\text{Min } \hat{Z}^t = \alpha^{t-1} Z^t = \alpha^{t-1}\left[\sum_{i\epsilon I'} \sum_{j\epsilon J''} c_{ij} x_{ij}^t + \sum_{i\epsilon I'} k_i^t N\right]$$

(13)
$$\sum_{j\epsilon J''} x_{ij}^t = a_i^0 \qquad \text{for } i\epsilon I',$$

(14) $$\sum_{i \in I'} x^t_{ij} = b^t_j \qquad \text{for } j \in J'',$$

(15) $$x^t_{i,\, n+2} \leqslant N \qquad \text{for } i \in I', \text{ and}$$

(16) $$x^t_{ij} \geqslant 0 \qquad \text{for } i \in I',\, j \in J''$$

where

$$x^t_{i,\, n+1} = s^t_i \qquad \text{for } i \in I',$$

$$x^t_{i,\, n+2} = N - q^t_i \qquad \text{for } i \in I',$$

$$a^0_i = q^0_i + N, \qquad c_{i,\, n+1} = h_i, \qquad c_{i,\, n+2} = -k'_i \qquad \text{for } i \in I', \qquad b^t_j = \sum_{\tau=0} r^\tau_j \qquad \text{for } j \in J'',$$

$$\sum_{\tau=0}^{t} r^\tau_{n+2} = (m+1)N + \sum_{i \in I'} q^0_i - \sum_{j \in J'} \sum_{\tau=0}^{t} r^\tau_j \text{ and } Z^t = \sum_{i \in I'} \sum_{j \in J''} c_{ij} x^t_{ij} + \sum_{i \in I'} k'_i N.$$

Problem $P3$ is likewise converted to problem $P4$, which consists of the series of subproblems $P4^t$; $t \in K$.

It should be noted that the coupling constraint (4), i.e.

$$q^t_i - q^{t-1}_i \geqslant 0 \qquad \text{for } i \in I,\, t \in K'$$

is equivalent to

(17) $$-x^t_{i,\, n+2} + x^{t-1}_{i,\, n+2} \geqslant 0 \qquad \text{for } i \in I,\, t \in K'.$$

We will now develop a solution procedure for solving $P4$ and then show that the optimal solution so obtained also satisfies (17). Hence the optimal solution to $P4$ so obtained is also an optimal solution to problem $P2$, the original problem we set out to solve. For ease of presentation of the solution procedure, we define the problem $P5^t$, which is

$$\text{Min } Z^t = \sum_{i \in I'} \sum_{j \in J''} c_{ij} x^t_{ij} + \sum_{i \in I'} k'_i N$$

over constraint set (13)–(16). The optimal solution to $P5^t$ is also optimal to $P4^t$, with $\hat{Z}^t = \alpha^{t-1} Z^t$.

3.2. Solution Strategy for Solving $P5^t$ (and $P4^t$)

Each $P5^t$ is a capacitated transportation problem with upper bounds only on the $(n+2)$ column. The objective function of a subproblem $P5^t$ is similar to the others. Further the requirement vector or rim condition [6, 7] of a $P5^t$ differs from $P5^{t-1}$ by amounts of r^t_j for $j \in J''$ (i.e., $b^t_j = b^{t-1}_j + \delta r^t_j$ for $j \in J''$ since $b^{t-1}_j = \sum_{\tau=0}^{t-1} r^\tau_j$ and $\delta = 1$).

With respect to the transportation tableau * of $P5^t$, it may be recalled that a cell (i, j) is an ordered index pair with $i \epsilon I', j \epsilon J''$. A basis B is a set of $(m+1)+(n+2)-1$ cells without cycles and such that there is at least one cell in each row and each column [5]. Let u_i and v_j be the dual variables associated with B. A basic optimum solution B^0 to $P5^0$ can be found by using any standard capacitated transportation algorithm [5] or more efficiently by using the procedure suggested in [8]. For each t, $P5^t$ can be solved parametrically starting from the optimal solution to $P5^{t-1}$ by utilizing the area rim operator $\delta \mathbf{R}_A$ (P) [6, 7]. This operator transforms the optimum solution of problem P into that of a problem of P^A whose data are the same as that of P except that $a_i^A = a_i + \delta\alpha_i$ and $b_j^A = b_j + \delta\beta_j$. For $P5^t$, $a_i^t = a_i^{t-1}$, $i \epsilon I'$ and $\beta_j = r_j^t$, $j \epsilon J''$. Hence letting $\alpha_i = 0$ for $i \epsilon I'$ and $\beta_j = r_j^t$ for $j \epsilon J''$, $P5^t$ can be solved by applying $1\mathbf{R}_A$ (P^{t-1}). The optimum solution to $P4^t$ is the same as the optimal solution to $P5^t$, with value of the objective function $\hat{Z}^t = \alpha^{t-1} Z^t$.

We provide below an algorithm to solve each $P4^t$ successively for $t = 1, 2, \ldots, T$ starting from B^0. The first phase of the algorithm determines an optimum basic solution to $P4^0$. The second phase applies the area rim operator to find the optimum solutions to $P4^t$, $t \epsilon K$. We denote by ψ the set of cells (i, j) which have their respective x_{ij} bounded from above. From (15),

(18) $$\Psi = \{(i, n+2) \text{ for } i \epsilon I'\}$$

Let $\overline{\Psi} = \{(i, j) \text{ for } i \epsilon I', j \epsilon J'\}$

Since only the cells in Ψ are bounded from above, we can somewhat simplify the statements of the algorithm in [6, 7] for applying $\delta \mathbf{R}_A(P)$. This simplified version is given below and illustrated with an example. The reader should refer to [6, 7] for the proofs involved in Phase II.

3.3 ALGORITHM 1: Parametric Procedure for Solving Problem P4.

Phase I: For finding a basic optimal solution to $P4^0$.

STEP 1: Solve $P5^0$ using any standard capacitated transportation procedure [5] or the procedure suggested in [8]. Record the B^0, $X^0 = \{x_{ij}^0\}$, u_i^0 for $i \epsilon I'$ and v_j^0 for $j \epsilon J''$. These are the optimal primal solutions to $P4^0$ as well. Also record Z^0. Set $\hat{Z}^0 = Z^0/\alpha$.

PHASE II: For applying $\delta \mathbf{R}_A(P)$ to solve $P4^t$, $t \epsilon K$.

STEP 2: Set $t = 1$, $k = 1$, $B_k^t = B^0$, $X_k^t = X^0$, $Z_k^t = Z^0$, $u_{i,k}^t = u_i^0$ for $i \epsilon I'$, $v_{j,k}^t = v_j^0$ for $j \epsilon J''$.

STEP 3: Set $\beta_j = r_j^t$ for $j \epsilon J''$. Find the modification matrix [6, 7], $Y = \{y_{ij}\}$ associated with β_j for $j \epsilon J''$ and B_k^t by solving the following equations:

(19) $$\sum_{j \epsilon J''} y_{ij} = 0 \qquad \text{for } i \epsilon I',$$

(20) $$\sum_{i \epsilon I'} y_{ij} = \beta_j \qquad \text{for } j \epsilon J'' \text{ and}$$

---

* The reader is assumed to be familiar with the usual definitions in the primal algorithm for the transportation problem [5]. For continuity the necessary definitions will be stated briefly when they are needed.

(21) $$y_{ij} = 0 \qquad \text{for } (i, j) \notin B^t_k.$$

Determine μ^t_k, the maximum extent to which the basis preserving area rim operator $\delta\mathbf{R}_A(P)$ can be applied using Equation (30) of [6], i.e.

(22) $$\mu^t_k = \text{Min}[(-x^t_{ij,k})/y_{ij}] \qquad \text{for } \{(i, j)\epsilon\bar{\Psi}|y_{ij} < 0\}$$

Record (r, s) as the cell at which the minimum of (22) is taken.

STEP 4: (i) If $\sum_{i=1}^{k} \mu^t_i \geq 1$, go to Step 4 (iii). Otherwise, go to Step 4 (ii).

(ii) (a) Set $b^t_j = b^{t-1}_j + \left(\sum_{i=1}^{k} \mu^t_i\right) \beta_j \qquad$ for $j\epsilon J''$,

$$Z^t_{k+1} = Z^t_k + \mu^t_k \left[\sum_{j\epsilon J''} \beta_j v_{j,}{}^t{}_k \right] \qquad \text{and } X^t_{k+1} = \{x_{ij,}{}^t{}_{k+1} = x_{ij,}{}^t{}_k + (\mu^t_k) y_{ij}\}$$

(b) Delete the cell (r, s) from B^t_k. If there are no basis cells in row r, define $I'_r = r$, $J''_r = \phi$, $I'_s = I' - \{r\}$, $J''_s = J''$. Otherwise imagine the remaining basis cells to be connected by horizontal and vertical lines whenever two of the cells lie in the same row or column. Define I'_r and J''_r to be those lines (rows and columns) that are connected to row r. Set $I'_s = I' - I'_r$, $J''_s = J'' - J''_r$.*

(c) Determine Δ and the cell (e, f) where the minimum is taken (Equation (58) of [7]) via:†

(23) $$\Delta = \text{Min} \begin{cases} (c_{ij} - u_{i,}{}^t{}_k - v_{j,}{}^t{}_k) & \text{for } (i, j)\epsilon[I'_s \times J''_r] \cap \Psi \\ (u_{i,}{}^t{}_k + v_{j,}{}^t{}_k - c_{ij,k}) & \text{for } (i, j)\epsilon[I'_r \times J''_s] \cap \mathrm{X} \end{cases}$$

Set $B^t_{k+1} = B^t_k - (r, s) + (e, f)$.

(d) Determine the duals $u_{i,}{}^t{}_{k+1}$ for $i\epsilon I'$ and $v_{j,}{}^t{}_{k+1}$ for $j\epsilon J''$ using Equation (60) of [7], i.e.

(24)
$$u^t_{i,k+1} = \begin{cases} u^t_{i,k} - \Delta & \text{for } i\epsilon I'_r \\ u^t_{i,k} & \text{for } i\epsilon I'_s \end{cases}$$

$$v^t_{j,k+1} = \begin{cases} v^t_{j,k} + \Delta & \text{for } j\epsilon J''_r \\ v^t_{j,k} & \text{for } j\epsilon J''_s \end{cases}$$

Set $k = k + 1$, go to Step (3).

---

*For a more rigorous definition of these sets, see [6].

†It will be proved later (Theorem 1) that $x_{i,}{}^t{}_{n+2,k}$ for $i\epsilon I'$ is nonincreasing as a function of the number of applications of Steps (3) and (4). Since in $P4^t$, $x_{i,}{}^t{}_{n+2}$ for $i\epsilon I'$ is the only set of variables with upper bounds, $x_{rs,}{}^t{}_{k+1} = 0$ always, and Equation (58) of [7] applies.

(iii) Set

$$\mu_k^t = 1 - \sum_{i=1}^{k-1} \mu_i^t, \qquad b_j^t = b_j^{t-1} + r_j^t \qquad \text{for } j \epsilon J'',$$

$$B_1^{t+1} = B_{k+1}^t = B_k^t, \qquad Z_1^{t+1} = Z_{k+1}^t = Z_k^t + \mu_k^t \left[\sum_{j \epsilon J''} \beta_j v_{j,k}^t \right],$$

$$u_{i,1}^{t+1} = u_{i,k+1}^t = u_{i,k}^t \qquad \text{for } i \epsilon I', \qquad v_{j,1}^{t+1} = v_{j,k+1}^t = v_{j,k}^t \qquad \text{for } j \epsilon J''.$$

Find $X_1^{t+1} = X_{k+1}^t = \{x_{ij,k+1} = x_{ij,k}^t + (\mu_k^t) y_{ij}\} \cdot X_{k+1}^t$ is a basic optimal solution to $P4^t$ with $\hat{Z}_{k+1}^t = \alpha^{t-1} Z_{k+1}^t$. Record them as X^t and $\hat{Z}^t$ respectively. Set $t = t+1$, $k = 1$. If $t > T$, go to Step 5. Otherwise go to Step 3.

STEP 5: X^t, $\hat{Z}^t$ for $t = 1, \ldots, T$ have been obtained. They are the optimal solution to $P4^t$. Stop.

3.4. A Numerical Example

Figure 1(a), (b), and (c) show a $P5^0$, $P5^1$ and $P5^2$ respectively for $m = n = 2$ (i.e. $P4$ for $m = n = 2$

$u_i \backslash v_j$	9	8	−1	−1	
0	20 (9)	40 (8)	2	N; N (−5)	60 + N
2	12	0 (10)	30 (1)	N; N (−2)	30 + N
1	M	M	0 (0)	N; N (0)	0 + N
	20	40	30	3N	

FIGURE 1(a). ($P5^0$)

9	8	2	N; −5	60 + N
12	10	1	N; −2	30 + N
M	M	0	N; 0	0 + N
20 + 20	40 + 40	30 + 60	3N − 120	

FIGURE 1(b). ($P5^1$)

9	8	2	N; −5	60 + N
12	10	1	N; −2	30 + N
M	M	0	N; 0	0 + N
20 + 20 +20	40 + 40 +20	30 + 60 +0	3N − 120 −40	

FIGURE 1(c). ($P5^2$)

FIGURE 1. Numerical example.

NOTE: $\hat{c}_{ij} = \alpha^{t-1} c_{ij}$, where $\hat{c}_{ij}$ and c_{ij} are the cost coefficients for $P4^t$ and $P5^t$ respectively.

LEGEND: U_{ij}; x_{ij} (c_{ij}) — U_{ij} is the upper bound on the cell (i, j)

(If not marked, $U_{ij} = \infty$.)

and $T=2$, with $\hat{c}_{ij}=\alpha^{t-1}c_{ij}$, where $\hat{c}_{ij}$ and c_{ij} are the cost coefficients for $P4^t$ and $P5^t$ respectively), and let $\alpha=0.8$. Note that

$$r_3^0=\sum_{i\epsilon I} q_i^0-\sum_{j\epsilon J} r_j^0=30 \qquad \text{and} \sum_{\tau=0}^{1} r_3^\tau=\sum_{\tau=0}^{2} r_3^\tau=\sum_{i\epsilon I} q_i^0=90.$$

Also,

$$r_4^0=3N+\sum_{i\epsilon I'} q_i^0-\sum_{j\epsilon J'} r_j^0=3N,$$

$$r_4^1=3N+\sum_{i\epsilon I'} q_i^0-\sum_{\tau=0}^{1}\sum_{j\epsilon J'} r_j^\tau=3N-120,$$

$$r_4^2=3N-120-40.$$

In Phase I of Algorithm 1, using the primal transportation algorithm [5], we obtain the optimum solution to $P5^0$ as shown in Figure 1(a). The circled cells denote optimum basis B^0 and the amounts x_{ij}^0 are shown in the northeast corners of the cells. $Z^0=530$, $\hat{Z}^0=(530/0.8)=662.5$.

In Phase II of Algorithm 1, we set $k=1$, $t=1$, $B_1^1=B^0$, $Z_1^1=Z^0$, $u_{i,1}^1=u_i^0$ for $i\epsilon I'$ and $v_{j,1}^1=v_j^0$ for $j\epsilon J''$. Figure 2(a) shows the modification matrix $Y=\{y_{ij}\}$ obtained by applying Step 3. $\mu_1^1=\operatorname{Min}\left[\frac{40}{20},\frac{30}{60}\right]=1/2$. Thus $(r,s)=(2,3)$.

In Step 4, since $\mu_1^1<1$, we go to Step 4(ii)(a). Now $b_j^1=b_j^0+\beta_j/2$ for $j\epsilon J''$, and $x_{ij,2}^1=x_{ij,1}^1+(1/2)y_{ij}$. These are shown in Figure 2(b). It can be verified that $Z_2^1=810$. From Step 4(ii)(b), $I_r'=\{2,1\}$, $I_s'=\{3\}$, $J_r''=\{2,1\}$, $J_s''=\{3,4\}$. From Step 4(ii)(c), we find that $\Delta=\operatorname{Min}[M-10, M-9, -1+5, 1+2]=3$ and $(e,f)=(2,4)$. $B_2^1=B_1^1-(2,3)+(2,4)$. Using (24) (Step 4(ii)(d)), we obtain the $u_{i,2}^1$ and $v_{j,2}^1$ for $i\epsilon I'$ and $j\epsilon J''$ as shown in Figure 2 (c). Applying Step (3), we obtain the modification matrix $Y=\{y_{ij}\}$ as shown in Figure 2(d). We have $\mu_2^1=3/2$, $(r,s)=(1,2)$. In Step 4 since $\sum_{i=1}^{2}\mu_i^1>1$ we go to Step 4(iii), where we set $\mu_2^1=1-1/2=1/2$, and $b_j^1=b_j^0+r_j^1$ for $j\epsilon J''$. B^1 and X^1 are as shown in Figure 2(e). It can be verified that $Z_3^1=1{,}180$. Hence $\hat{Z}^1=(0.8)^0(1{,}180)=1{,}180$. We now have the optimum solution to $P4^1$.

	20	−20			0
		60	−60		0
			120	−120	0
β_j	20	40	60	−120	

$Y=\{y_{ij}\}$

FIGURE 2(a)

$u^1_{i,1}$ \ $v^1_{j,1}$	9	8	−1	−1	
0	30 9	30 8	2	$\boxed{N}$ N −5	60 + N
2	12	30 10	1	$\boxed{N}$ N −2	30 + N
1	M	M	60 0	$\boxed{N}$ N − 60 0	0 + N
	20 + 10	40 + 20	30 + 30	3N − 60	

FIGURE 2(b)

$u^1_{i,2}$ \ $v^1_{j,2}$	12	11	−1	−1	
−3	30 9	30 8	2	$\boxed{N}$ N −5	60 + N
−1	12	30 10	1	$\boxed{N}$ N −2	30 + N
1	M	M	60 0	$\boxed{N}$ N − 60 0	0 + N
	20 + 10	40 + 20	30 + 30	3N − 60	

FIGURE 2(c)

	20	−20		
		60		−60
			60	−60
β_j	20	40	60	−120

$Y = \{y_{ij}\}$

FIGURE 2(d)

$u^1_{i,3}$ \ $v^1_{j,3}$	12	11	−1	−1	
−3	40 9	20 8	2	$\boxed{N}$ −5	60 + N
−1	12	60 10	1	$\boxed{N}$ N − 30 −2	30 + N
1	M	M	90 0	$\boxed{N}$ N − 90 0	0 + N
	20 + 20	40 + 40	30 + 60	3N − 120	

FIGURE 2(e)

FIGURE 2. Optimum solution to $P4^1$.

Proceeding along similar steps, starting from B^1, we obtain the optimum solution to $P4^2$, with B^2 and X^2 as shown in Figure 3(a). $\hat{Z}^2 = (0.8)[1{,}680] = 1{,}344$. The optimum solution to $P4$ is X^1 and X^2 with $\hat{Z} = \hat{Z}^1 + \hat{Z}^2 = 2{,}524$.

$u^2_{i,1}$ \ $v^2_{j,1}$	12	11	-1	-1	
-3	60 (9)	0 (8)	2	N -5	$60+N$
-1	12	100 (10)	1	N $N-70$ (-2)	$30+N$
1	M	M	90 (0)	N $N-90$ (0)	$0+N$
	$40+20$	$80+20$	$90+0$	$3N-120$ -40	

FIGURE 3(a)

u^*_i \ v^*_j	0	0	-0.8	-0.8
0	0	0	0	-1.6
0	0	0	0	-0.8
$+0.8$	0	0	0	0

FIGURE 3(b)
(γ_{ij}'s are marked in the northeast corner of the cells.)

FIGURE 3. Optimum solution to $P4^2$.

3.5. Equivalence Between *P4* and *P2*

We note below a key property of Algorithm 1.

THEOREM 1: (i) The values of $x^t_{i,n+2,k}$ obtained on the application of Steps (3) and (4) form a nonincreasing function of k for a particular t.

(ii) The values of $x^t_{i,n+2}$ in X^t form a nonincreasing function of t.

PROOF: (i) For a particular t, by Step 4(ii), $x^t_{ij,k+1} = x^t_{ij,k} + (\mu^t_k) y_{ij}$ where $Y = \{y_{ij}\}$ is computed in Step 3 for t and k. We shall show that $y_{i,n+2} \leq 0$ for $i \in I'$. Assume the contrary, i.e.

$$\exists \text{ a } y_{i_0,n+2} > 0, \qquad i_0 \in I'. \tag{25}$$

From (19) $\sum_{j \in J''} y_{i_0,j} = 0$, hence

$$\exists\, y_{i_0,j_1} < 0, \qquad j_1 \in J'' - \{n+2\}. \tag{26}$$

But from (20) $\sum_{i \in I'} y_{i,j_1} = \beta_j = r^t_{j_1} \geq 0$ hence,

$$\exists\, y_{i_1,j_1} > 0, \qquad i_1 \in I' - \{i_0\}. \tag{27}$$

But from (19) $\sum_{j \in J''} y_{i_1,j} = 0$ hence,

$$\exists\, y_{i_1,j_2} < 0, \qquad j_2 \in J'' - \{j_1\}. \tag{28}$$

If $j_2 = \{n+2\}$, the cycle $\{(i_0, n+2), (i_0, j_1), (i_1, j_1), (i_1, n+2)\}$ has been detected in B_k^t which is a contradiction.

If $j_2 \neq \{n+2\}$, since $r_{j_2}^t \geq 0$

(29) $$\exists\, y_{i_2, j_2} > 0, \qquad i_2 \epsilon I' - \{i_0\} - \{i_1\}.$$

($i_2 \neq i_0$, since if $i_2 = i_0$, the cycle $\{(i_0, j_1), (i_1, j_1), (i_1, j_2), (i_0, j_2)\}$ has been detected by (26)-(29), contradicting the fact that B_k^t is a basis.)

We are then back to (28) with a

$$y_{i_2, j_3} < 0, \qquad j_3 \epsilon J'' - \{j_1\} - \{j_2\}.$$

(Again $j_3 \neq j_1$ since if $j_3 = j_1$ a cycle has been detected in B_k^t.)

Since I' and J'' are finite, inductively, we must terminate with the existence of a

(30) $$y_{i_p, n+2} < 0.$$

Conditions (25)-(30) imply the existence of a cycle in B_k^t, which is a contradiction. Now, $\mu_k^t \geq 0$, hence we have $x_{i,n+2,k+1}^t \leq x_{i,n+2,k}^t$ for $i \epsilon I'$.

(ii) For a particular t and k, by Step 4(iii)

$$x_{i,n+2,1}^{t+1} = x_{i,n+2}^{t} \qquad \text{for } i \epsilon I'$$

By Theorem 1(i), we have

$$x_{i,n+2}^{t+1} = x_{i,n+2,l}^{t+1} \leq x_{i,n+2,1}^{t+1} = x_{i,n+2}^{t} \text{ where } l \geq 1.$$ Q.E.D.

REMARK 1: By Theorem 1(ii), Algorithm 1 provides an optimal solution to $P4$ with $x_{i,n+2}^t$ a nonincreasing function of t, for $i \epsilon I'$. Consequently, the coupling constraint (17) (which is equivalent to (4)) is satisfied. Therefore, the optimal solution provided by Algorithm 1 is also an optimal solution to the linear capacity expansion problem $P2$.

Continuing on the numerical example in section 3.4, we find that $q_1^0 = q_1^1 = q_1^2 = 0$, $q_2^0 = 0$, $q_2^1 = 30$, $q_2^2 = 70$. Hence the optimal solution X^1, X^2 is also an optimal solution to the problem $P2$, formed by adding the coupling constraint $q_i^t - q_i^{t-1} \geq 0$ for $t = 2$, $i = 1, 2$ to the numerical example given in Figure 1.

REMARK 2: Algorithm 1 provides an optimal solution to $P2$, by solving parametrically (in terms of the rim conditions) $T+1$ transportation subproblems $P4^t$, $t = 0, 1, \ldots, T$. Since the computational times for such large subproblems (e.g. 25×500) are only a few seconds, Algorithm 1 can provide an optimal solution to fairly large size problems $P2$ (e.g. $I = 25$, $J = 500$ and $T = 10$) in a matter of seconds. However, as stated before, a problem $P2$ os such size may be computationally unwieldy for linear programming codes.

REMARK 3: In converting problem $P1$ to $P2$ we assumed that $g_i = k_i$ for $i \epsilon I$. We show below that this assumption can be somewhat relaxed. If $k_i \neq g_i$ then the expression $\alpha^{T-1}[\alpha(k_i - g_i)]q_i^T$ will have

to be added on to the objective function (7). In terms of the problem $P5$ this affects only $P5^T$ in which the costs $c^T_{i,n+2}$ change to $\bar{c}^T_{i,n+2}$ where

$$\bar{c}^T_{i,n+2} = c^T_{i,n+2} - \alpha(k_i - g_i).$$

In terms of the operator theory of parametric programming [6, 7] this is referred to as the area cost operator $\delta\mathbf{C}_A(P)$ which transforms the optimum solution of P to that of P^A with the same data except $c^A_{ij} = c_{ij} + \delta\gamma_{ij}$ for $i \epsilon I$ and $j \epsilon J$. In our problem $\delta = 1$, $\gamma_{i,n+2} = -\alpha(k_i - g_i)$ for $i \epsilon I$ and $\gamma_{ij} = 0$ elsewhere. Let the maximum extent to which $\delta C_A(P)$ can be applied without altering the current basis structure be μ_A. Then if $\mu_A \geqslant 1$ then B^T is also optimal to this changed set of costs. However if $\mu_A < 1$ then further application of the operator involves basis changes and the solution so obtained for $\delta = 1$ may not satisfy the constraints (17). However, for a reasonably long planning horizon, even in the latter case, the solution obtained with the assumption $k_i = g_i$ may be expected to be nearly optimal to $P1$.

Continuing on the example from section 3.4 we have $k_1 = 5/0.2 = 25$ and $k_2 = 10$. Let $g_1 = 23$ and $g_2 = 9$. Hence we have $\gamma_{1,4} = -(0.8)(2) = -1.6$, $\gamma_{2,4} = -0.8$ and $\gamma_{ij} = 0$ for the remaining cells. Applying Theorem 8 of [6], we obtain the u^*_i for $i \epsilon I'$ and v^*_j for $j \epsilon J''$ as shown in Figure 3(b). The maximum extent to which the $\delta C_A(P)$ can be applied is (see Equation (40) of [6])

$$\mu^A = \text{Min}[(M-1-12)/0.8,\ (M-1-11)/0.8] = (M-13)/0.8$$

Hence the (primal) optimal solution as given in Figures 2(e) and 3(a) remains optimal to $P1$ with $g_1 = 23$ and $g_2 = 9$ as well. In fact for this example, the optimal solution remains optimal to $P1$ for $0 \leqslant g_1 \leqslant k_1$ and $0 \leqslant g_2 \leqslant k_2$.

4.0. OPTIMALITY PROPERTIES AND INFINITE PLANNING HORIZON

4.1. Properties of the Optimal Solution

In this section we study some properties of the optimal solution to $P2$ as provided by Algorithm 1.

LEMMA 2: In the optimal solution to $P2$ as provided by Algorithm 1, $s^t_i (\equiv x^t_{i,n+1}) = 0$ for all $t \epsilon \{\tau, \tau+1, \ldots, T\}$ if $q^\tau_i > 0$ $(\equiv x^\tau_{i,n+2} < N)$ for any $\tau \epsilon K$ and $i \epsilon I'$.

PROOF: Assume the contrary, i.e. $s^t_i > 0$ for a $t \epsilon \{\tau, \tau+1, \ldots, T\}$.

By Theorem 1(ii) $q^t_i \geqslant q^\tau_i > 0$. The value of the objective function to subproblem $P4^t$ can be reduced by decreasing s^t_i ($\equiv$ decreasing $x^t_{i,n+1}$) and decreasing q^t_i ($\equiv$ increasing $x^t_{i,n+2}$), contradicting the fact that $x^t_{i,n+1}$ and $x^t_{i,n+2}$ belong to an optimal solution of $P4^t$.

The physical interpretation of Lemma 2 is intuitively meaningful. It says that with linear costs, there exists an optimal solution to $P2$ such that if there is a capacity addition in a period for some region, then that region's full capacity will be utilized for that period as well as all future periods.

LEMMA 3: If q^0_i, and $\sum_{\tau=0}^{t} r^\tau_j$ are integers for $i \epsilon I$, $j \epsilon J$ and $t \epsilon \{0\} + K$, the optimal solution to $P2$ as provided by Algorithm 1 are also integers.

PROOF: $\sum_{\tau=0}^{t} r^\tau_{n+1}$ and $\sum_{\tau=0}^{t} r^\tau_{n+2}$ for $t \epsilon \{0\} + K$ involve additions and subtractions of q^0_i and $\sum_{\tau=0}^{t} r^\tau_j$, $i \epsilon I$, $j \epsilon J$ and $t \epsilon \{0\} + K$. Also by definition, $q^0_{m+1} = 0$. Since N can also be assumed to be an integer, the rim conditions of $P4^t$ (see (13)–(15) i.e., a^0_i and b^t_j for $i \epsilon I$ and $j \epsilon J''$, $t \epsilon \{0\} + K$) are all integers. The

constraint matrix of each subproblem $P4^t$, $t \in \{0\}+K$ is totally unimodular. Application of Algorithm 1, therefore, produces integral X^t for $t \in K$. By Remark 1, this integral solution is optimal to $P2$.

Lemma 3 is interesting because the constraint matrix of $P2$ is itself *not* unimodular. In the physical context of the problem, integral optimal solution may be important, since it may be meaningless to add a nonintegral number of capacity units. In the multiperiod assignment problem mentioned in section 1, this may be particularly important.

LEMMA 4: (i) Assume $r_j^t > 0$ for $j \in J'$.* For some t and k, if at Step 3 of Algorithm 1, the modification matrix $Y=\{y_{ij}\}$ obtaining by solving (19)–(21) has $y_{ij} \geq 0$, for $i \in I'$, $j \in J'$ then $\mu^t = L$ (a large positive number).

(ii) The $B^t = B_{k+1}^t$ derived from the subsequent application of Step 4(iii) remains an optimal basis for all subproblems $P4^\tau$, $\tau \geq t$.

PROOF: We have $y_{ij} \geq 0$ for $i \in I'$, $j \in J'$. Using equation (22) we have

$$\mu_k^t = \mathrm{Min}(-x_{i,n+2,k}^t)/y_{i,n+2} \text{ for } \{(i,\ j) \in \overline{\Psi} \mid y_{ij} < 0\}$$

Since the set of minimization is empty, $\mu_k^t = L$ (a large positive number)

(iii) The $Y=\{y_{ij}\}$ satisfy (19)–(21). Since $\mu_k^t = L$, from Step (3) we go into Step 4(ii) where $B^t = B_{k+1}^t$ is obtained. At the next application of Step 3, we can obtain a new modification matrix $\hat{Y}=\{\hat{y}_{ij}\}$ satisfying (19)–(21) by defining $\hat{y}_{ij}=(r_j^{t+1}/r_j^t)y_{ij}$ for $i \in I'$, $j \in J'$ and $\hat{y}_{i,n+2} = -\sum_{j \in J'} \hat{y}_{ij}$ for $i \in I'$. Since $r_j^t > 0$, $y_{ij} \geq 0$† and $r_j^{t+1} \geq 0$ for $i \in I'$, $j \in J'$, we have $\hat{y}_{ij} \geq 0$ for $i \in I'$, $j \in J'$. $\hat{Y}$ thus satisfies the condition of Lemma 4(i) and $\mu_1^{t+1} > 1$. Hence $B^{t+1} = B^t$. The argument can be repeated to show that $B^\tau = B^t$ for $\tau \geq t$.

Lemma 4 provides a sufficient condition at which Algorithm 1 can be terminated, since when this condition has been attained at time t (say), an optimal solution to all $P4^\tau$, $\tau \geq t$ can be readily obtained. An optimal solution to $P2$ is given by $B^1, B^2, \ldots, B^t, B^{t+1}=B^t, \ldots, B^T=B^t$, and $X^1, X^2, \ldots, X^t$, $X^{t+1}=X^t+Y^{t+1}, X^{t+2}=X^{t+1}+Y^{t+2}, \ldots, X^T=X^{T-1}+Y^T$. ($Y^{t+1}$ is the modification matrix obtained by solving (19)–(21) with $\beta_j = r_j^{t+1}$ for $j \in J''$.)

4.2. Infinite Horizon Case

We now show that Algorithm 1 can be applied to solve $P1$ with an infinite planning horizon. Obviously, in the infinite horizon case, the assumption regarding the resale value of the capacity at the end of the planning horizon is not needed. Infinite planning models have received interest e.g. [3, 4] as it is felt that any terminal condition imposed on a planning model is arbitrary.

We assume that for periods prior to T, the demands in each market can be estimated with greater accuracy. Hence the growth rates for these periods may be any nonnegative rate. At the end of the planning period T, the demand in each market j is estimated to grow at the rate β_j per period.** After

---

*If $r_j^t = 0$ for any $j \in J''$, let $r_j^t = \epsilon$, where ϵ is a very small positive number.

†It can be shown easily that since $y_{ij} \geq 0$ for $i \in I'$, $j \in J'$, there can exist one and only one $y_{ij} > 0$ for each $j \in J'$.

**From the definitions given in Section 3.1, we have $\beta_{n+1}=0$ and $\beta_{n+2} = -\sum_{j \in J} \beta_j$.

solving $P4^t$ for t up to T, the remaining problem in an infinite horizon model consists of the subproblems $P4^t$ for $t > T$ where

$$b_j^t = \sum_{\tau=0}^{t} r_j^\tau = \sum_{\tau=0}^{T} r_j^\tau + \beta_j(t-T) \qquad \text{for } t > T \text{ and } j \epsilon J''.$$

It is obvious that there exists a basic feasible solution to any subproblem $P4^t$, for $t \geqslant 0$. Each $P4^t$ is an equivalent problem to the others except for the requirement vector [5], which is a nondecreasing function of t. By a well known result in parametric programming [5, p. 149], there exists an optimum basis B^τ to $P4^\tau$ which remains optimal to all $P4^t$, $t \geqslant \tau$.

Algorithm 1 can be applied to solve $P4^t$, $t \geqslant 0$ until the modification matrix Y (as computed in Step 3) satisfies the condition of Lemma 4(i). Assume this is achieved at $t = t_1$. A basis B^{t_1} has been obtained which remains optimal to all $P4^t$, $t \geqslant t_1$. An optimal solution to $P4^t$, $t > t_1$ is given by $X^t = X^{t-1} + Y^t$. By Lemma 4, $B^1, B^2, \ldots, B^{t_1}, B^{t_1+1} = B^{t_1}, \ldots B^\infty = B^{t_1}$ is an optimal solution to $P1$ for the infinite horizon case.

REFERENCES

[1] Erlenkotter, D., "Preinvestment Planning for Capacity Expansion—A Multi-Location Dynamic Model", Ph. D. Dissertation, Graduate School of Business, Stanford University, Stanford, California, 1969.

[2] Fong, C. O. and M. R. Rao, "Capacity Expansion with Two Producing Regions and Concave Costs", Management Science, *22*, 331–339 (1975).

[3] Hopkins, D. S. P., "Infinite-Horizon Optimality in an Equipment Replacement and Capacity Expansion Model", Management Science, *18*, 145–156 (1971).

[4] Manne, A. S. (Ed.), *Investment for Capacity Expansion: Size, Location and Time Phasing* (MIT Press, Cambridge, Massachusetts, 1967).

[5] Simonnard, M., *Linear Programming*, (Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1966).

[6] Srinivasan, V., and G. L. Thompson, "An Operator Theory of Parametric Programming for the Transportation Problem—I", Nav. Res. Log. Quart., *19*, 205–225 (1972).

[7] Srinivasan, V., and G. L. Thompson, "An Operator Theory of Parametric Programming for the Transportation Problem—II", Nav. Res. Log. Quart., *19*, 227–252 (1972).

[8] Srinivasan, V., and G. L. Thompson, "Determining Optimal Growth Paths in Logistics Operations", Nav. Res. Log. Quart., *19*, 575–599 (1972).

AN ALGORITHM FOR SOLVING GENERAL FRACTIONAL INTERVAL PROGRAMMING PROBLEMS*

A. Charnes

University of Texas
Austin, Texas

Daniel Granot and Frieda Granot

University of British Columbia
Vancouver, B.C., Canada

ABSTRACT

A Linear Fractional Interval Programming problem (FIP) is the problem of extremizing a linear fractional function subject to two-sided linear inequality constraints.

In this paper we develop an algorithm for solving (FIP) problems. We first apply the Charnes and Cooper transformation on (FIP) and then, by exploiting the special structure of the pair of (LP) problems derived, the algorithm produces an optimal solution to (FIP) in a finite number of iterations.

1. INTRODUCTION

Problems of maximizing a linear fractional objective function subject to two-sided linear inequality constraints were termed in the literature as fractional interval programming problems, denoted by (FIP). Their general formulation is

$$\begin{aligned} &\text{Max } (c^T x + c_0)/(d^T x + d_0) \\ &\text{subject to} \quad b^- \leq Ax \leq b^+ \end{aligned}$$

We remark that for a suitable choice of the vectors b^- and b^+ the constraints set of (FIP) is sufficiently general to cover all bounded polyhedral sets.

Problems with linear fractional objective function arise, e.g. in attrition games [14], Markovian replacement problems [11, 15], reduction of integer programs to knapsak problems [4], the cutting stock problem [13], in primal dual approaches to decomposition procedures [1, 16].

The linear interval constraints arise in problems of capital budgeting, blending and mixing problems, production planning problems and more, see e.g. [3, 19].

---

*This research was partially supported by NRC Grant number A-4024 and by ONR Contracts N00014-67-A-0126-0008 and N00014-67-A-0126-0009 with the Center for Cybernetic Studies, The University of Texas.

A complete analysis and an explicit solution for (FIP) when the coefficient matrix A is of full row rank was first obtained by Charnes & Cooper [7], see also [5, 10]. A finite algorithm for solving the general (FIP) problem directly, i.e. without resorting to the transformation of (FIP) to an equivalent linear programming problem, was constructed in [9].

In this paper another finite primal algorithm for solving (FIP) is developed. In contrast with the algorithm developed in [9], we first apply here Charnes and Cooper's transformation [6] on (FIP) to reduce it to a pair of (LP) problems. The crucial observation is that for a fixed value of the additional variable t, introduced by that transformation, each one of the pair of (LP) problems is an Interval Programming problem (IP) which is significantly smaller in size. This last feature is fully exploit in our algorithm where instead of solving the pair of (LP) problems directly, we solve parametrically for t the associated (IP) problems using the method developed in [8]. We start with a feasible value of t for which an optimal solution to the associated (IP) problem is generated. The values of t and the corresponding optimal solutions are then modified until, after finitely many iterations, an optimal solution to (FIP) is produced.

Thus, by exploiting the special structure of the interval constraints and by using the primal algorithm for (IP) problems [8] as the main device, we are able to efficiently solve (FIP) problems.

The algorithm to be described here together with that developed in [9] are based on two of the major approaches suggested for solving fractional programs—those of Charnes and Cooper [6] and Martos [17], respectively, and they both utilize the special structure of the interval constraints while generating an optimal solution to (FIP).

2. PRELIMINARY RESULTS

Consider again the (FIP) problem:

$$\text{Max}\left\{\frac{c^Tx+c_0}{d^Tx+d_0}=\frac{C(x)}{D(x)}\right\} \tag{1}$$

$$\text{s.t.}\quad x\epsilon S=\{b^-\leqslant Ax\leqslant b^+\} \tag{2}$$

where c^T, c_0, d^T, d_0, b^-, b^+ and A are given. In the sequel we shall assume that S is bounded and

$$\frac{c^Tx+c_0}{d^Tx+d_0}\neq \text{constant} \tag{3}$$

on S.

LEMMA 1 [7]: A feasible (FIP) is unbounded if either $c\notin N(A)^\perp$, or $d\notin N(A)^\perp$, where $N(A)$ is the null space of A.

LEMMA 2 [9]: Let (FIP) be given with $c\epsilon N(A)^\perp$, $d\epsilon N(A)^\perp$ and $A\epsilon R_r^{m\times n}$ where $R_r^{m\times n}$ is the set of all $m\times n$ matrices with rank r. Let $D\epsilon R_r^{r\times n}$ satisfy

$$R(D^T)=R(A^T)$$

Then:

(a) $AD^T\epsilon R_r^{m\times r}$.

(b) (FIP) is equivalent to a full column rank (FIP) with coefficient matrix AD^T and cost function

$$\frac{c^T D^T y + c_0}{d^T D^T y + d_0}.$$

(c) If feasible, the optimal solutions of (FIP) are $D^T y^* + N(A)$ where y^* is the set of optimal solutions to the equivalent problem.

PROOF: Since $R(A^T) = N(A)^\perp$ and since $c \in N(A)^\perp$, $d \in N(A)^\perp$ it follows that every optimal solution to (FIP) is of the form

$$x^* + N(A)$$

where x^* is an optimal solution to

$$\text{Max } \frac{c^T x + c_0}{d^T x + d_0}$$

$$\text{(4)} \qquad \text{s.t.} \quad b^- \leq Ax \leq b^+$$

$$x \in R(A^T)$$

But, since $R(A^T) = R(D^T)$, $x \in R(A^T)$ can be equivalently written as:

$$x = D^T y, \qquad y \in R^r$$

Substituting $x = D^T y$ in (4) results in the equivalent problem, which completes the proof.

3. AN ALGORITHM FOR SOLVING (FIP)

Consider again the (FIP) problem

$$\text{(1)} \qquad \text{Max} \left\{ \frac{c^T x + c_0}{d^T x + d_0} = \frac{C(x)}{D(x)} \right\}$$

$$\text{(2)} \qquad \text{s.t.} \quad b^- \leq Ax \leq b^+$$

and assume that (FIP) is feasible, $c \perp N(A)$, $d \perp N(A)$ (see Lemma 1) and A is of full column rank representation (see Lemma 2).

Following Charnes and Cooper's transformation [6], we multiply $C(x)$ and $D(x)$ by t, $t > 0$, and restrict $t \cdot D(x)$ to be equal to 1 (or -1, for negative values of the denominator). Substituting

$$\text{(5)} \qquad tx = z$$

in (1), (2) results with the following pair of (LP) problems, equivalent to (FIP) and denoted by (ELP1) ((ELP2))

(6) $$\text{Max } \{c^T \cdot z + c_0 \cdot t\} \qquad (\text{Max } (-c^T \cdot z - c_0 \cdot t))$$

(7) $$\text{s.t.} \quad A \cdot z \leqslant b^+ \cdot t$$

(8) $$A \cdot z \geqslant b^- \cdot t$$

(9) $$d^T \cdot z = 1 - d_0 \cdot t \qquad (d^T \cdot z = -1 - d_0 \cdot t)$$

(10) $$0 \leqslant t \leqslant M$$

where M is a sufficiently large number. As it was noted in [6], (ELP1) and (ELP2) differ from each other by only a change in sign in the functional and in one constraint.

We observe that the constraints of (ELP1) ((ELP2)) are of special structure. Indeed, for any fixed value of t, the $2m$ constraints (7), (8) are equivalent to m interval constraints of the form

(11) $$b^- \cdot t \leqslant A \cdot z \leqslant b^+ \cdot t.$$

Thus, for any fixed feasible value of t (ELP1) ((ELP2)) can be reduced to an equivalent (IP) problem denoted by (EIP1) ((EIP2)) of the form

(12) $$\text{Max } c^T \cdot z + c_0 \cdot t \qquad (\text{Max} - c^T \cdot z - c_0 \cdot t)$$

(13) $$\text{s.t.} \quad b^- \cdot t \leqslant A \cdot z \leqslant b^+ \cdot t$$

(14) $$1 - d_0 \cdot t \leqslant d^T z \leqslant 1 - d_0 \cdot t \qquad (-1 - d_0 \cdot t \leqslant d^T z \leqslant -1 - d_0 \cdot t).$$

Let x^1 be a feasible solution to (FIP), not necessarily an extreme point.

REMARK 1: For a real world problem, a feasible solution to (2) might sometimes be at hand from the available data on the problem. If, however, a feasible solution to (FIP) is not available then $x = 0$ can be chosen as a feasible start whenever $b^- \leqslant b^+$. Otherwise, assuming $b^+ \geqslant 0$, we can solve the following (IP) problem

$$\text{Max} - M \cdot e^T U$$

$$\text{s.t.} \quad b^- \leqslant Ax + IU < b^+$$

$$0 \leqslant IU \leqslant b^+$$

where M is sufficiently large, in order to produce a feasible solution to (FIP), if such a solution exists. For a more detailed discussion see [8].

Clearly, if $D(x^1) > 0$ $(D(x^1) < 0)$ then

(15) $$t_1 = 1/D(x^1),\ z^1 = t_1 \cdot x^1 \quad . \quad (t_1 = -1/D(x^1),\ z^1 = t_1 \cdot x^1)$$

is a feasible solution to (ELP1) ((ELP2)).

In the following, we shall present a method for solving (ELP1) i.e., the case that $D(x^1) > 0$. Exactly the same method will be applied for solving (ELP2) when $D(x^1) < 0$.

The main idea underlying our method is to take advantage of the special structure of (ELP1) while attempting to generate an optimal solution to (FIP). Thus for $t = t_1$ we first apply the primal algorithm for (IP) problems, introduced in [8], in order to produce an optimal solution $z(t_1)$, to (EIP1).

Let B be the optimal basis generated by the primal algorithm for (IP) and N the completion of the rows of $\binom{A}{d^T}$ to B. Let $b^+(t) = \binom{b^+ \cdot t}{1-d_0 \cdot t}$ and $b^-(t) = \binom{b^- \cdot t}{1-d_0 \cdot t}$, and denote by $b_B^-(t)$, $b_N^-(t)$, $b_B^+(t)$, $b_N^+(t)$ the partitions of the vectors $b^-(t)$, $b^+(t)$ which correspond to the partition of A to B and N, respectively. Substituting

$$y = Bz \tag{16}$$

in (EIP1) and rearranging the order of the constraints results with the following equivalent problem

$$\text{Max } \{c^T \cdot B^{-1}y + c_0 \cdot t\} \tag{17}$$

$$\text{s.t.} \quad b_B^-(t) \leq y \leq b_B^+(t) \tag{18}$$

$$b_N^-(t) \leq NB^{-1}y \leq b_N^+(t) \tag{19}$$

Since B is an optimal basis, y^1 given by

$$y_i^1 = \begin{cases} b_B^+(t_1)_i & \text{if } c^T B_i^{-1} \geq 0 \\ b_B^-(t_1)_i & \text{if } c^T B_i^{-1} < 0 \end{cases} \qquad (i = 1, \ldots, n) \tag{20}$$

is an optimal solution to (17), (18), (19) for $t = t_1$, and

$$z(t_1) = B^{-1} \cdot t_1 \tag{21}$$

is an optimal solution to (EIP1ff for $t = t_1$. See also [2], [8].

Clearly, $t = t_1$ is not necessarily the optimal value of t in (ELP1). In the following we shall generate the optimal value of t, t^{opt}, and the optimal solution $z(t^{\text{opt}})$ to (EIP1) for $t = t^{\text{opt}}$.

Let

$$b_B(t)_i = \begin{cases} b_B^+(t)_i & \cdot \text{ if } c^T B_i^{-1} \geq 0 \\ b_B^-(t)_i & \text{if } c^T B_i^{-1} < 0 \end{cases} \qquad (i = 1, \ldots, n) \tag{22}$$

Treating t as a variable and substituting

$$y_i = b_B(t)_i \; (i = 1, \ldots, n) \tag{23}$$

in (17), (18), (19) results with the following single variable constrained maximization problem

$$\operatorname{Max}\left\{\sum_{i=1}^{n}(c^TB^{-1})_i b_B(t)_i + c_0 \cdot t\right\} \equiv \operatorname{Max}\{\alpha_B \cdot t + \beta_B\} \tag{24}$$

$$\text{s.t.} \quad b_N^-(t) \leqslant NB^{-1}b_B(t) \leqslant b_N^+(t). \tag{25}$$

Since $t = t_1$ is a feasible solution to (25), then, if $\alpha_B = 0$ we can immediately conclude that $(z(t_1), t_1)$ is an optimal solution to (ELP1). If, however, $\alpha_B \neq 0$ we need to consider the following three exhaustive cases:

(i) $\alpha_B > 0$ and the value of t can be increased until M without violating any of the constraints in (25), then $C(x)/D(x)$ is not bounded over S, see [6].

(ii) $\alpha_B > 0(\alpha_B < 0)$ and the value of t can be increased (decreased) from t_1 to t_2, where $t_2 > t_1(t_2 < t_1)$ is the largest (smallest) value of t which does not violate the constraints in (25).

(iii) Any ϵ increase (decrease) in the value of t when $\alpha_B > 0(\alpha_B < 0)$ violates at least one of the constraints in (25).

Thus, whenever $\alpha_B = 0$ or case (i) occurs we terminate with the appropriate conclusions.

Assume therefore that we encountered case (ii), $\alpha_B > 0$ and that we had increased the value of t until t_2. We shall attempt to vary the value of t from t_2 so as to improve the value of the objective function. Exactly the same method is applied if case (ii), $\alpha_B < 0$ occurs or when case (iii) occurs, for $t = t_1$.

Since for each j, $b_j^+(t)$, $b_j^-(t)$ are linear functions of t, they can be written as

$$b_j^+(t) = b_j^+ \cdot t + \gamma_j^+, \qquad b_j^-(t) = b_j^- \cdot t + \gamma_j^- \tag{26}$$

For simplicity, let the nonbasic constraint satisfied as equality at t_2 be

$$a_1y_1(t_2) + \ldots + a_n \cdot y_n(t_2) = b(t_2)_{n+1} = b_{n+1} \cdot t_2 + \gamma_{n+1} \tag{27}$$

where $b(t_2)_{n+1} = b^+(t_2)_{n+1}$ if the nonbasic constraint reached its upper bound as t was increased to t_2 or $b(t_2)_{n+1} = b^-(t_2)_{n+1}$ if it reached its lower bound.

We shall refer to nonbasic constraints satisfied as equalities at $t = t_2$ as critical nonbasic constraints, and to $t = t_2$ as a critical value of t.

REMARK 2: We shall assume in the sequel that there exists only one critical non-basic constraint at $t = t_2$ and at any other critical value of t. This assumption can always be made, since if not, a perturbation, essentially equivalent to that introduced in [8] for the linear interval programming problem can be performed in order to secure this property. The perturbed problem is obtained from the original problem by replacing the vectors b^+ and b^- by the perturbed vectors $b^+(\epsilon)$ and $b^-(\epsilon)$ where

$$b_i(\epsilon) = b_i^+ + \epsilon^i \qquad b_i^-(\epsilon) = b_i^- - \epsilon^i \qquad (i = 1, \ldots, m)$$

and ϵ is sufficiently small and positive.

Let

$T = \{t;\ t \text{ is feasible to (ELP1)}\}$

$v(t)$ —The optimal value of the objective function in (ELP1) for $t \epsilon T$.

LEMMA 3: T is convex and $v(t)$ is concave over T.

PROOF: The convexity of T is clear. Next, let $t_1, t_2 \epsilon T$ and let $\lambda, \bar{\lambda}$ be non-negative scalars such that $\lambda + \bar{\lambda} = 1$. Then,

(28)
$$\begin{aligned} v(\lambda t_1 + \bar{\lambda} t_2) &= \operatorname*{Max}_{z_1, z_2} \{c^\top [\lambda z_1 + \bar{\lambda} z_2] + c_0[\lambda t_1 + \bar{\lambda} t_2] \\ &\qquad \text{s.t.} \quad b^-[\lambda t_1 + \bar{\lambda} t_2] \leqslant A[\lambda z_1 + \bar{\lambda} z_2] \leqslant b^+[\lambda t_1 + \bar{\lambda} t_2]\} \\ &\geqslant \operatorname*{Max}_{z_1, z_2} \{c^\top [\lambda z_1 + \bar{\lambda} z_2] + c_0[\lambda t_1 + \bar{\lambda} \cdot t_2] \\ &\qquad \text{s.t.} \quad b^-(t_1) \leqslant A z_1 \leqslant b^+(t_1), \quad b^-(t_2) \leqslant A z_2 \leqslant b^+(t_2)\} \\ &= \lambda \cdot \operatorname{Max} \{c^\top z \\ &\qquad \text{s.t.} \quad b^-(t_1) \leqslant A z \leqslant b^+(t_1)\} \\ &+ \bar{\lambda} \cdot \operatorname{Max} \{c^\top z \\ &\qquad \text{s.t.} \quad b^-(t_2) \leqslant A z \leqslant b^+(t_2)\} = \lambda \cdot v(t_1) + \bar{\lambda} v(t_2) \end{aligned}$$

which completes the proof; see also [[12] Lemma 1].

Since it can be shown that (23) is optimal to (17), (18), (19) for $t_1 \leqslant t \leqslant t_2$, then, as a corollary to Lemma 3, we have the optimal value of t in (ELP1) is greater than or equal to t_2.

The feasible solution $(y(t_2), t_2) = (y_1(t_2), \ldots, y_n(t_2), t_2)$ is an extreme point for (18), (19) in R^{n+1}. In the following, we shall generate an adjacent edge to $(y(t_2), t_2)$ along which, by increasing the value of t we shall improve the value of the objective function (if such an incident edge exists).

Suppose we form a new basis by removing δ_j, the jth basic constraint, and inserting the critical nonbasic constraint. Then, upon substituting the new set of values of the y_j's (as it was done in order to obtain (24), (25)), we can calculate the slopes of the objective function and the constraints, as functions of t, in the new basis.

Denote by

B=The current basis.

B_j=B/ The jth basic constraint $\cup$ the critical nonbasic constraint, where / denotes deletion and $\cup$ denotes union.

α_B=The slope of the objective function in the current basis B.

α_{Bj}=The slope of the objective function in the basis B_j.

S_{n+1}=The slope of the critical nonbasic constraint in the basis B.

S_i=The slope of the jth basic constraint in the new basis B_j.

Then it is easy to verify that

(29)
$$\alpha_{Bj} = \alpha_B + \frac{(c^T B^{-1})_j}{a_j} [b_{n+1} - S_{n+1}]$$

(30)
$$S_j = (b_B)_j + \frac{1}{a_j} [b_{n+1} - S_{n+1}].$$

We conclude therefore that if $b_B^-(t_2)_j < b_B^+(t_2)_j$ and either

$$(31) \qquad (b_B)_j = (b_B^-)_j,\ \alpha_{Bj} > 0,\ S_j \geq (b_B^-)_j$$

or

$$(32) \qquad (b_B)_j = (b_B^+)_j,\ \alpha_{Bj} > 0,\ S_j \leq (b_B^+)_j$$

then we can increase the value of t, without violating any of the nonbasic constraints, and thus improve the value of the objective function. If in the jth basic constraint, $b_B^-(t_2)_j = b_B^+(t_2)_j$ we shall attempt to remove that constraint from the basis only if $S_j = (b_B^+)_j = (b_B^-)_j$ and $\alpha_{B_j} > 0$.

Thus, if possible, the value of t will be increased until its next critical value, i.e. the largest value that t can be assigned without violating any of the nonbasic constraints.

THEOREM 1: If for each j, $(j = 1, \ldots, n)$ neither (31) nor (32) are satisfied then

$$(33) \qquad (z(t_2),\ t_2)$$

is an optimal solution to (ELP1), where $z(t_2) = B^{-1}y(t_2)$.

PROOF: Replacing a basic constraint with the critical nonbasic constraint and attempting to increase the value of the objective function through an increase in t, amounts to examining the possibility of improving the value of the objective function by moving along an incident edge to $(y(t_2), t_2)$. Since any such attempt failed, and since $y(t_2)$ is an optimal solution to (17), (18), (19) for $t = t_2$ we conclude that none such incident edge exists, which implies that $(y(t_2), t_2)$ is an optimal solution to (17), (18), (19), and thus $(B^{-1} \cdot y(t_2), t_2)$ is an optimal solution to (ELP1).

As we noted earlier, we maintain an optimal tableau while increasing (or decreasing) the value of t from t_1 until t_2. Explicitly, for each $t \in [t_1, t_2]$, (23) is an optimal solution to (17), (18), (19). However, this property does not generally hold as we move, according to our criteria, from one extreme point to an adjacent extreme point by varying the value of t. Therefore, whenever we reach a critical value of t we need to check the optimality of the current basic solution. If it is optimal we try to improve the value of the objective function by moving along one of the incident edges to the current basic solution. If, however, the solution is not optimal, we first apply the primal algorithm for (IP) problems in order to generate an optimal solution for that critical value of t.

Algorithm A for solving (FIP) is conveniently summarized in the following.

ALGORITHM A.

Step 1: Generate a feasible solution x^1 to (FIP)—from which obtain by (15) a feasible solution (z^1, t_1) either to (ELP1) and then set $i = 1$ or to (ELP2) and then set $i = 2$.

Step 2. Fix the value of t at its current value and solve (ELPi) by the primal algorithm for (IP) problems. Can the value of t be modified (i.e. increased or decreased) without violating any of the constraints, which will result with an improved value for the objective function?

No, $\alpha_B = 0$—terminate with an optimal solution to (ELPi), go to step 5.

No, $\alpha_B \neq 0$—go to step 4.

Yes—go to step 3.

Step 3. Can the value of t be increased to M without violating any of the nonbasic constraints?

Yes—terminate, with the conclusion that $C(x)/D(x)$ is not bounded over S.

No—Increase (or decrease) the value of t, whereas the values of the y_i's are modified accordingly, as it was done in (23), until an additional nonbasic constraint is satisfied as equality, i.e. until the next critical value of t.

Do the y_i's satisfy the optimality criterion for the current value of t?

Yes—go to step 4.

No—go to step 2.

Step 4. Check whether by removing a basic constraint and inserting a critical nonbasic constraint the value of the objective function can be improved by modifying the value of t.

If impossible—terminate with an optimal solution to (ELPi), go to step 5.

If possible—perform the appropriate basis change and go to step 3.

Step 5. Are the optimal solutions to (ELP1) and (ELP2) at hand?

No—find a feasible solution (z^1, t_1) to (ELPj) $j \epsilon \{1, 2\}$ $j \neq i$, set $i = j$, go to step 2.

Yes—Let $(z(t_1^{\text{opt}}), t_1^{\text{opt}})$ and $(z(t_2^{\text{opt}}), t_2^{\text{opt}})$ be optimal solutions to (ELP1), (ELP2), respectively.

Then

$$
x^{\text{opt}} = \begin{cases} \dfrac{z(t_1^{\text{opt}})}{t_1^{\text{opt}}} & \text{if } c^T[z(t_1^{\text{opt}})] + c_0 \cdot t_1^{\text{opt}} \geqslant c^T[z(t_2^{\text{opt}})] + c_0 \cdot t_2^{\text{opt}} \\[2em] \dfrac{z(t_2^{\text{opt}})}{t_2^{\text{opt}}} & \text{if } c^T[z(t_2^{\text{opt}})] + c_0 \cdot t_2^{\text{opt}} \geqslant c^T[z(t_1^{\text{opt}})] + c_0 \cdot t_1^{\text{opt}} \end{cases} \tag{34}
$$

is an optimal solution to (FIP).

4. OPTIMALITY AND CONVERGENCE

The finiteness of Algorithm A stems from the finiteness of the primal algorithm for (IP) and from the fact that while we move from one extreme point of (ELPi) $i = 1, 2$ to an adjacent extreme point we strictly increase (in the perturbed problem if necessary, see remark 2) the value of the objective function. Since the number of extreme points is finite so is Algorithm A.

REMARK 3: Let B be an optimal basis and $y(t_k)$ an optimal solution for (17), (18), (19) for $t = t_k$. Assume that at $t = t_k$ the first nonbasic constraint is a critical constraint. Thus, (17), (18), (19) can be written as:

$$
\begin{aligned}
&\text{Max } (c^T B^{-1})_1 y_1 + \ . \ . \ . + (c^T B^{-1})_n y_n + c_0 t \\
&\text{s.t.} \quad b_B^-(t)_1 \leqslant y_1 \leqslant b_B^+(t)_1 \\
&\qquad\quad \vdots \\
&\qquad\quad b_B^-(t)_n \leqslant y_n \leqslant b_B^+(t)_n
\end{aligned}
$$

$$b^-(t)_{n+1} \leqslant a_1 y_1 + \ . \ . \ . \ + a_n y_n \leqslant b^+(t)_{n+1}$$
$$b^-_{\tilde{N}}(t) \leqslant \tilde{N} B^{-1} y \leqslant b^+_{\tilde{N}}(t)$$

where $\tilde{N}$ are the remaining nonbasic rows.

If we choose to replace the critical nonbasic constraint, which is satisfied as equality on its upper bound (lower bound) with the δ_i basic constraint for which

(i) $a_i \neq 0$

(ii) $\dfrac{(c^T B^{-1})_i}{a_i} \geqslant 0 \qquad \left(\dfrac{(c^T B^{-1})_i}{a_i} \leqslant 0\right)$

(iii) $\dfrac{(c^T B^{-1})_i}{a_i} \leq \dfrac{(c^T B^{-1})_j}{a_j} \qquad \left(\dfrac{(c^T B^{-1})_i}{a_i} \geq \dfrac{(c^T B^{-1})_j}{a_j}\right), a_j \neq 0,\ j \neq i$

Then, the tableau obtained after performing the proper basis change is also optimal, i.e. the new y_i's satisfy the optimality criterion (20).

The above basis change will be performed only if it will be possible to modify the value of t in the new basis and thus to improve the value of the objective function. In this case we would not have to solve an (IP) problem at the next critical value of t (see step 2 in Algorithm A).

However, we remark that it might be impossible to construct a new optimal basis which will include the critical nonbasic constraint, and even if possible, we might not be able to improve the value of the objective function in the new basis, by modifying the value of t, in which case we will operate according to Algorithm A.

EXAMPLE [9]: Solve the (FIP) problem:

$$\text{Max } \frac{3x_1 - x_3 + 4}{2x_2}$$

$$\begin{aligned} \text{s.t.} \quad -1 \leqslant\ & x_1 \qquad + x_3 \leqslant 2 \\ 0 \leqslant\ & x_1 + x_2 \qquad \leqslant 7 \\ 1 \leqslant\ & \qquad x_2 \qquad \leqslant 5 \\ 0 \leqslant\ & \qquad\qquad x_3 \leqslant 1 \end{aligned}$$

Clearly, since $x_2 \geqslant 1$ (ELP2) is not feasible and thus we need only to solve (ELP1). A feasible solution to (FIP) is $x^1 = (0, 1, 0)$ for which $D(x^1) > 0$. From (15) we get $z^1 = (0, 1/2, 0)$, $t_1 = 1/2$ as a feasible solution to (ELP1). Problem (EIP1) is of the form:

(EIP1):

$$\text{Max } 3z_1 - z_3 + 4t$$

$$\text{s.t.} \quad -t \leqslant z_1 \qquad + z_3 \leqslant 2t$$

$$0 \leq z_1 + z_2 \leq 7t$$

$$t \leq z_2 \leq 5t$$

$$0 \leq z_3 \leq t$$

$$1/2 \leq z_2 \leq 1/2$$

$$\text{Let } B = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad B^{-1} = \begin{pmatrix} 1 & 0 & -1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad N = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \end{pmatrix}$$

Then,

$$c^T B^{-1} = (3, 0, -4).$$

Substituting $y = Bz$ in (EIP1) results with

$$\text{Max } 3y_1 - 4y_3 + 4t$$

$$\text{s.t.} \quad -t \leq y_1 \leq 2t$$

$$1/2 \leq y_2 \leq 1/2$$

$$0 \leq y_3 \leq t$$

$$0 \leq y_1 + y_2 - y_3 \leq 7t$$

$$t \leq y_2 \leq 5t$$

An optimal solution y is $y = (2t, 1/2, 0)$. Substituting y as a function of t in the objective function and the nonbasic constraints results with

$$\text{Max } 6t + 4t$$

$$\text{s.t.} \quad 0 \leq 2t + 1/2 \leq 7t$$

$$t \leq 1/2 \leq 5t$$

Since $t_1 = 1/2$, the second nonbasic constraint is critical. However, only the second basic constraint can be replaced by the critical constraint and for this basic constraint we have $b_2^- = b^+$. According to our method this constraint can not be removed since $b_2^- = b_2^+ \neq S_2$. Thus $t^{\text{opt}} = 1/2$ is the optimal value for t and

$$z^{\text{opt}} = B^{-1}y = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ 1/2 \\ 0 \end{pmatrix} = \begin{pmatrix} 1 \\ 1/2 \\ 0 \end{pmatrix}$$

The optimal solution for (FIP) is

$$x^{\text{opt}} = \frac{z^{\text{opt}}}{t^{\text{opt}}} = \begin{pmatrix} 2 \\ 1 \\ 0 \end{pmatrix}.$$

REFERENCES

[1] Bell, E. J., "Primal-Dual Decomposition Programming", Ph.D. Thesis, Operations Research Center, Univ. of California at Berkeley, Report ORC 65–92, (1965).

[2] Ben-Israel, A. and A. Charnes, "An Explicit Solution of a Special Class of Linear Programming Problems" Operations Research *16*, 1166–1175 (1968).

[3] Ben-Israel, A., A. Charnes, A. P. Hurter and P. D. Robers, "On the Explicit Solution of a Special Class of Linear Economic Models", Operations Research *18*, 462–470 (1970).

[4] Bradley, G., "Transformation of Integer Programs to Knapsack Problems", Discrete Mathematics, *1*, 29–45 (1971).

[5] Bühler, W., "A Note on Fractional Interval Programming", Technesche Hochschule Aachen, West Germany, (Jan. 1973).

[6] Charnes, A. and W. W. Cooper, "Programming with Linear Fractional Functionals", Nav. Res. Log. Quart., *9*, 181–186 (1962).

[7] Charnes, A. and W. W. Cooper, "An Explicit General Solution in Linear Fractional Programming", Nav. Res. Log. Quart., *20*, 449–467 (1973).

[8] Charnes, A., D. Granot and F. Granot, "A Primal Algorithm for Interval Linear Programming Problems", Department of Mathematics, Dalhousie University, (July 1974).

[9] Charnes, A., D. Granot, and F. Granot, "On Solving Linear Fractional Interval Programming Problems", Department of Mathematics, Dalhousie University, (August 1974).

[10] Charnes, A., D. Granot and F. Granot, "A Note on Explicit Solution In Linear Fractional Programming", Nav. Res. Log. Quart., *23*, 161–167 (1976).

[11] Derman, C., "On Sequential Decisions and Markov Chains", Management Science, *9*, 16–24 (1962).

[12] Geoffrion, A. M., "Duality in Nonlinear Programming: A Simplified Application—Oriented Development", SIAM Review, *13*, 1–37 (1971).

[13] Gilmore P. C. and R. E. Gomory, "A Linear Programming Approach to the Cutting Stock Problem—Part II", Operations Research, *11*, 863–888 (1963).

[14] Isbell, J. R. and W. H. Marlow, "Attrition Games", Nav. Res. Log. Quart. *3*, 71–93 (1956).

[15] Klein, M., "Inspection—Maintenance—Replacement Schedules under Markovian Deterioration", Management Science, *9*, 25–32 (1962).

[16] Lasdon, L. S., *Optimization Theory for Large Systems; Chapters II and IV* (The MacMillan Company, Collier—MacMillan Limited, London, 1970).

[17] Martos, B., "Hyperbolic Programming", translated by A. and V. Whinston, Nav. Res. Log. Quart., *11*, 135–155 (1964).

[18] Martos, B., "The Direct Power of Adjacent Vertex Programming Methods", Management Science, *12*, 241–252 (1965).

[19] Robers, P. D., "Interval Linear Programming", Ph.D. Thesis submitted to Northwestern University (1968).

SOLVING FIXED CHARGE NETWORK PROBLEMS WITH GROUP THEORY-BASED PENALTIES

R. L. Rardin and V. E. Unger

School of Industrial and Systems Engineering
Georgia Institute of Technology

ABSTRACT

Many well-known transportation, communication, and facilities location problems in operations research can be formulated as fixed charge network problems, i.e. as minimum cost flow problems on a capacitated network in one commodity where some arcs have both fixed and variable costs. One approach to solving such problems is to use group theoretic concepts from the theory of integer programming to provide bounds for a branch-and-bound procedure. This paper presents such a group-theory based algorithm for exact solution of fixed charge network problems which exploits the special structures of network problems. Computational results are reported for problems with as many as 100 fixed charge arcs.

1. INTRODUCTION

The fixed charge network problem (FCNP) can be formulated as

$$\min c_1^T x_1 + c_2^T x_2 + c_s^T s \tag{1}$$

$$\text{s.t.} \quad E_1 x_1 + E_2 x_2 = 0 \tag{2}$$

$$-Ix_1 + Iy - Is = 0 \tag{3}$$

$$u_1 \geqslant y \geqslant 0 \tag{4}$$

$$u_1 \geqslant x_1 \geqslant 0 \tag{5}$$

$$u_1 \geqslant s \geqslant 0 \tag{6}$$

$$u_2 \geqslant x_2 \geqslant l_2 \tag{7}$$

$$y \equiv 0 \bmod u_1, \tag{8}$$

where c_1, c_s, s, u_1, x_1 and y are n_1-vectors; c_2, l_2, u_2 and x_2 are n_2-vectors; b is an m-vector; E_1 is $m \times n_1$; E_2 is $m \times n_2$; $c_s > 0$; and (E_1, E_2) is the node-arc incidence matrix of a direct network, i.e. a matrix with columns which correspond to arcs of the network and consist entirely of zeros except

for a -1 in the row corresponding to the origin node of the arc and a $+1$ in the row corresponding to the destination node of the arc. The problem is a general minimum cost flow problem in one commodity with the added features that all variables are bounded, and that a subset of the flow variables (the x_{1j}) have a fixed charge cost structure. Thus each x_{1j} has both a variable charge $v_j x_{1j}$ and a positive fixed charge f_j assessed whenever, $x_{1j} > 0$. To obtain the formulation FCNP, prorata amounts $c_{sj} = f_j/u_{1j}$ are calculated, the cost c_{1j} is defined by $c_{1j} = c_{sj} + v_j$, and constraints on the slack vector s are arranged so that the full amount f_j will be assessed when $x_{1j} > 0$.

In this paper the solution of fixed charge network problems in the form of FCNP will be pursued by using the group theoretic approach to integer programming developed by Gomory, Johnson and others [2, 3, 4, 5, 6, 9, 10, 11] to provide bounds for a branch-and-bound approach. Results in [15] for general fixed charge problems (i.e. problems like FCNP where no special structure of E_1 and E_2 is assumed) are specialized to the network case. An algorithm using the results is then outlined and computational results presented.

2. NOTATION

In order to effectively present observations about FCNP, some notational conventions will be required. For the convenience of readers, these conventions are summarized below.

1. Sets of rows from a matrix (or vector) will be denoted by enclosing the matrix in brackets and indicating the limiting row numbers. For example,

$$[M]_k^1 = \text{the submatrix consisting of rows 1 through } k \text{ of } M.$$

When only a single row of a matrix is required, the convention will be simplified by dropping the redundant superscript, and if no confusion will result the brackets will also be omitted leaving, for example,

$$x_{1j} = \text{the } j\text{th row or component of the vector } x_1$$

2. All references to optimal solutions, bases, and tableaux for various linear programs will be with respect to bases of the well-known bounded simplex procedure. When it is desired to speak of the part of a solution vector, cost vector, bound vector or matrix associated with the basic variables, nonbasic variables, etc., the usual rearrangement of rows and columns will be assumed, and identifying superscripts will be attached to submatrices. Specifically, the superscript B will denote the basic part of the matrix, N the nonbasic part, U the part with nonbasic variables at their upper bounds, and L the part with nonbasic variables at their lower bounds.

3. A bar over the name of a problem will denote the continuous relaxation of the problem, i.e. the same problem with any congruence constraints relaxed. Elements of the optimal solution, optimal simplex tableau, etc. for such a continuous relaxation will be similarly denoted by bars over the names of the elements.*

* Throughout the paper the usual simplification of referring to an optimal solution as if it were unique is observed. There may, of course, be many basic solutions which satisfy simplex optimality criteria. However, the given results hold for any such solution, and the only possible effect of alternative solutions is to make some penalties zero because the adjusted costs of corresponding nonbasic variables are zero.

4. The function nu (ν) will be used to denote the value of an optimal solution to the problem given as its argument.

3. RESULTS FOR GENERAL FIXED CHARGE PROBLEMS

The analysis of [15] developed a number of structural results for general fixed charge problems which will be exploited in the algorithm of this paper. The next few sections briefly outline the most important of these results. Though the results are stated in terms of the notation of FCNP, they apply to any bounded fixed charge problem.

3.1. The Continuous Relaxation

At each iteration of a group theory-based branch-and-bound solution procedure the continuous relaxation $\overline{\text{FCNP}}$ of a FCNP must be solved. In [15] it was shown that all elements of any optimal basic solution to $\overline{\text{FCNP}}$ could be constructed from the solution to the reduced network problem (RNP)

$$\min c_1^T x_1 + c_2^T x_2 \tag{9}$$

$$\text{s.t.} \quad E_1 x_1 + E_2 x_2 = 0 \tag{10}$$

$$u_1 \geqslant x_1 \geqslant 0 \tag{11}$$

$$u_2 \geqslant x_2 \geqslant l_2 \tag{12}$$

If $\{x_1^B, x_2^B\}$ forms an optimal basis for RNP, $\{x_1^B, x_2^B, y\}$ is an optimal basis for $\overline{\text{FCNP}}$. The corresponding optimal solution is given by

$$\begin{aligned} y = x_1 &= \bar{x}_1 \\ x_2 &= \bar{x}_2 \\ s &= 0, \end{aligned}$$

where $\{\bar{x}_1, \bar{x}_2\}$ is the optimal solution to RNP.

3.2. Penalty Subproblems

An equivalent form of any linear mixed-integer program can always be obtained by solving the continuous relaxation of the problem and rewriting the problem in terms of perturbations from the values of the nonbasic variables in the optimal continuous solution. The optimal simplex tableau is used to represent changes on basic variables in terms of perturbations in the nonbasics. It is shown in [15] that such an equivalent form for FCNP is given by

$$\min (\bar{c}_1^N)^T \begin{pmatrix} \Delta x_1^L \\ -\Delta x_1^U \end{pmatrix} + (\bar{c}_2^N)^T \begin{pmatrix} \Delta x_2^L \\ -\Delta x_2^U \end{pmatrix} + c_s^T \Delta s + \nu(\overline{\text{FCNP}}) \tag{13}$$

$$\text{s.t.} \quad \begin{pmatrix} [\bar{E}_1^N] \, {}^{1}_{k^B} \\ -I \end{pmatrix} \begin{pmatrix} \Delta x_1^L \\ -\Delta x_1^U \end{pmatrix} + \begin{pmatrix} [\bar{E}_2^N] \, {}^{1}_{k^B} \\ 0 \end{pmatrix} \begin{pmatrix} \Delta x_2^L \\ -\Delta x_2^U \end{pmatrix} - \Delta s \equiv \bar{y} \bmod u_1 \tag{14}$$

$$\Delta s,\ \Delta x_1^L,\ \Delta x_1^U,\ \Delta x_2^L,\ \Delta x_2^U \geqslant 0 \tag{15}$$

$$\begin{pmatrix} u_1^B \\ u_2^B \end{pmatrix} \geqslant \begin{pmatrix} \bar{x}_1^B \\ \bar{x}_2^B \end{pmatrix} - \begin{pmatrix} \bar{E}_1^N \end{pmatrix} \begin{pmatrix} \Delta x_1^L \\ -\Delta x_1^U \end{pmatrix} - \begin{pmatrix} \bar{E}_2^N \end{pmatrix} \begin{pmatrix} \Delta x_2^L \\ -\Delta x_2^U \end{pmatrix} \geqslant \begin{pmatrix} 0 \\ l_2^B \end{pmatrix} \tag{16}$$

$$\begin{pmatrix} u_1^B \\ u_1^N \end{pmatrix} \geqslant \bar{y} - \begin{pmatrix} [\bar{E}_1^N]\,{}^1_{k^B} \\ -I \end{pmatrix} \begin{pmatrix} \Delta x_1^L \\ -\Delta x_1^U \end{pmatrix} - \begin{pmatrix} [\bar{E}_2^N]\,{}^1_{k^B} \\ 0 \end{pmatrix} \begin{pmatrix} \Delta x_2^L \\ -\Delta x_2^U \end{pmatrix} + \Delta s \geqslant 0 \tag{17}$$

$$u_1 \geqslant \Delta s \tag{18}$$

$$u_1^L \geqq \Delta x_1^L,\ u_1^U \geqslant \Delta x_1^U \tag{19}$$

$$(u_2^L - l_2^L) \geqslant \Delta x_2^L,\ (u_2^U - l_2^U) \geqslant \Delta x_2^U \tag{20}$$

where
$$\begin{aligned} \Delta x_1^L &= x_1^L - 0 & \Delta x_2^U &= u_2^U - x_2^U \\ \Delta x_1^U &= u_1^U - x_1^U & \Delta s &= s - 0. \\ \Delta x_2^L &= x_2^L - l_2^L & k^B &= \text{the dimension of } x_1^B, \end{aligned}$$

and $\bar{c}_1^N, \bar{c}_2^N, \bar{E}_1^N, \bar{E}_2^N$ are taken from the optimal simplex tableau for RNP. The relaxation of this equivalent form consisting of (13), (14), and (15) is the group problem of Gomory and Johnson (denoted GP (FCNP)).

GP(FCNP) is generally difficult to solve. Thus a number of relaxations were developed in [15] to produce penalties for a branch-and-bound procedure. Relaxations of GP(FCNP) where only the rows of (14) with numbers in an index set Γ are enforced are denoted GP(Γ). When a GP(Γ) is further constrained to satisfy (18), (19), and (20), it is referred to as a *bounded group problem* and denoted BGP(Γ). Finally, when a GP(Γ) is constrained to satisfy all rows i of (17) for $i \in \Gamma$, it is referred to as an *either-or problem* and denoted EOP(Γ). The name for the latter problem derives from the fact that EOP(Γ) is GP(Γ) with "$= \bar{y}_i \bmod u_{1i}$" replaced by "$= \bar{y}_i$ or $u_{1i} - \bar{y}_i$" in row i of (14) for each $i \in \Gamma$.

Observe that all these subproblems can be constructed directly from the optimal simplex tableau for RNP. Moreover, the bounds obtained from solving the subproblems have the following obvious relationships:

$$\begin{aligned} &\nu\,[\mathrm{GP}(\Gamma)] \geqslant \nu[(\overline{\mathrm{FCNP}})] \\ &\nu\,[\mathrm{EOP}(\Gamma)] \geqslant \nu[\mathrm{GP}(\Gamma)] \\ &\nu\,[\mathrm{BGP}(\Gamma)] \geqslant \nu[\mathrm{GP}(\Gamma)] \end{aligned}$$

$$\left.\begin{aligned} &\nu\,[\mathrm{GP}(\Gamma)] \geqslant \nu[\mathrm{GP}(\Gamma')] \\ &\nu\,[\mathrm{EOP}(\Gamma)] \geqslant \nu[\mathrm{EOP}(\Gamma')] \\ &\nu\,[\mathrm{BGP}(\Gamma)] \geqslant \nu[\mathrm{BGP}(\Gamma')] \end{aligned}\right\} \qquad \Gamma' \subset \Gamma.$$

4. THE NETWORK CASE

The algorithm to be presented in this paper employs a branch-and-bound approach where the above subproblems are used to provide penalties. In such an approach certain of the congruence-constrainted variables y_j are fixed at either 0 or u_{1j}, and the continuous linear program obtained by relaxing congruence constraints on the other y_j variables is solved. From the optimal simplex tableau for this linear program a series of penalty problems is constructed to obtain lower bounds on how much the value of the continuous optimal linear programming solution would have to be increased to produce an optimal solution for the original fixed charge problem. These bounds or penalities are then used to reduce the set of remaining possibilities for optimal solutions to FCNP and choose a new branching variable, i.e. a new variable to fix so that enumeration can continue.

4.1. Solution of the Continuous Relaxation

The above discussion of properties of general fixed charge problems highlighted the importance of the reduced problem RNP in the various steps of such a branch-and-bound procedure. In the network case RNP is by definition a minimum cost flow problem in one commodity. Thus, the various steps in the branch-and-bound procedure involving RNP can be simplified by exploiting the network structure.

One of the most powerful sets of theory for minimum cost flow problems is the graph theoretic approach development by Johnson [8] and Langley [13] and others. In the terminology of this approach a *graph* is a collection of arcs and nodes associated with some network; a *cycle* is a connected set of two-ended arcs of the graph which touches nodes in such a way that every node is touched by exactly two arcs; a *tree* is a connected set of two-ended arcs which contains no cycles; and a *forest* is a set of trees. A forest is said to *span* a graph if each node is touched by exactly one tree. If a one-ended arc is added to each tree so that the number of nodes is equal to the number of arcs, the one-ended arc is called a *root*, and the tree is said to be a *rooted tree*. A collection of such rooted trees is a *rooted forest*, and a rooted forest which spans a network is a *rooted spanning forest*.

In terms of these definitions, the fundamental result on which the graph theoretic approach to network flow problems is based can be stated as follows:

4.1.1. Theorem (see [8]).

The arcs associated with any basis of a network flow problem like RNP form a rooted spanning forest for the network.

Define the node of a rooted tree touched by the root as the *base* of the tree. Then the importance of Theorem 4.1.1 derives from the fact that by systematically searching from the base of each tree in the spanning forest associated with a basis for a problem like RNP, it is possible to reach all basic arcs and all nodes without cycling. In particular define the direction *up* in a tree as away from the base of a tree, and *down* as toward the base. Similarly, nodes and arcs will be said to be *above* a given node or arc in a tree if they can be reached by preceeding up the tree from the given node or arc. Then by maintaining arc flows, dual multipliers, and the following labels, it is possible to easily perform all simplex operations necessary to solve RNP. (See for example [14] for the details of a simplex procedure.)

4.1.2. Definition.

The *basis label* of a node w in a network problem like RNP is $[\delta(w), \mu(w), \gamma(w), \alpha(w)]$ where

$\delta(w)$ = the number of the node directly below w in the basis forest (0 if w is the base of a tree).

$\mu(w)$ = the number of a node directly above w in the basis forest (0 if no such node exists).

$\gamma(w)$ = the number of a node z such that $\delta(w) = \delta(z)$ and $\gamma(\hat{z}) \neq z$ for all $\hat{z} \neq w$ satisfying $\delta(w) = \delta(\hat{z})$ (0 if no such node exists).

$\alpha(w)$ = the number of the arc connecting w and $\delta(w)$.

The components of the basis label are referred to as the *down node*, *up node*, *right node*, and *down arc*, respectively.

Figure 1 provides one of several such sets of basis labels for a sample basis forest. For example, the down node label of node 4 is 7 because 7 is the next closer node to the base of the forest, and the down arc label of node 4 is 17 because arc 17 connects nodes 4 and 7. The chain of nodes immediately above node 3 begins with node 2, which is the up node from 3, and proceeds right to node 8.

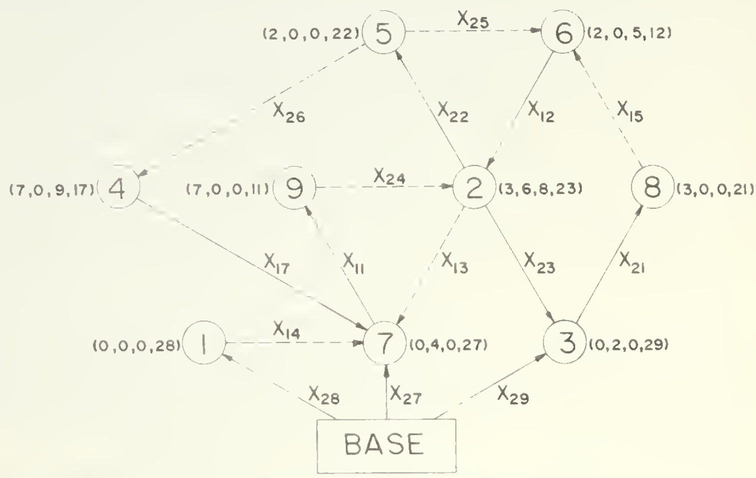

FIGURE 1. Basis labels for an RNP example.

4.2. Generating Constraint Matrices for Subproblems

The constraint matrices of the various subproblems defined in section 3.2 consist of rows drawn from the constraints (14) and (17). Careful study of the expressions for (14) and (17) will, in turn show that all nontrivial rows of these constraints (i.e. rows 1, 2, . . ., k^B) are extracted directly from the rows of the optimal RNP simplex tableau which correspond to basic components of x_1. Thus, the essential problem in generating constraint matrices for the penalty subproblems is to generate the rows of the updated simplex tableaux for RNP which correspond to basic components of x_1.

Suppose now that RNP has be solved by a graph-theoretic simplex procedure. Since the simplex tableau corresponding to the optimal labels has never been explicitly calculated, a procedure for generating nontrivial rows of the penalty subproblems from the labels is required.

The usual approach for generating updated simplex tableaux from initial tableaux is to premultiply the original tableaux by a basis inverse. In the case of an original tableau which is a node-arc incidence matrix, the process amounts at most to taking the difference of two columns in the basis inverse because there is at most one $+1$ and one -1 in each column of the original tableau.

From the above observations it follows that constraint matrices for the various penalty subproblems can easily be generated if an optimal basis inverse for RNP can be constructed from an optimal set of basis labels. The following theorem from the theory of the graph-theoretic approach to network problems shows that this can be easily accomplished by using the labels to trace the nodes of the network above a particular arc in the optimal basis forest.

4.2.1. Theorem

The row of the basis inverse corresponding to any basic arc of RNP will have +1 entries for all nodes above the arc in the basis forest if the arc is directed away from the base of its tree, and −1 entries for all nodes above the arc in the basis forest if the arc is directed toward the base of its tree. All nodes not above the arc in the basis forest will have 0 entries.

PROOF: See [13] p. 55.

Before turning to an example, one additional observation can be made. Recall that only the rows of the optimal tableau for RNP corresponding to basic components of x_1 are required to generate subproblems. Thus the corresponding rows of the basis inverse are the only ones required, and the following definition will lead to a further simplification.

4.2.2. Definition

A *macro-node* of a basis forest for RNP is a single node used to replace any maximal set of ordinary nodes in the forest of RNP which are connected by a tree of basic arcs drawn entirely from the vector x_2.

The effect of grouping nodes of a network into macro-nodes is to collapse the optimal basis forest for RNP into a tree consisting entirely of arcs with fixed charges (i.e. components of x_1). Figure 2 illustrates this reduction for the example of Figure 1.

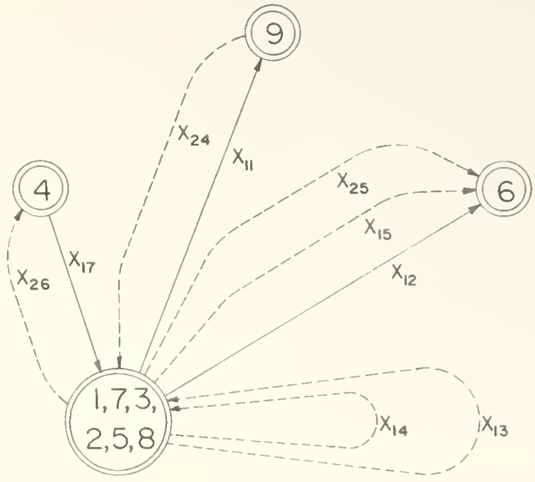

FIGURE 2. The macro-node tree for the example RNP.

Associated with this reduction in the complexity of the basis forest for RNP is a simplification in calculation of the basis inverse. The following theorem states the results.

4.2.3. Theorem

The columns of a basis inverse for RNP corresponding to any two nodes w and z which are part of the same macro-node will have equal entries in all rows associated with components of x_1^B.

PROOF: Let x_{1j} be any basic element of x_1 and Ψ be the set of nodes above x_{1j} in the basis forest for RNP. For any two nodes w and z which are members of the same macro-node, the unique path between the nodes in the basis forest for RNP consists entirely of arcs corresponding to elements of x_2. Thus x_{1j} is not a part of that path and either

$$w \in \Psi \quad \text{and} \quad z \in \Psi$$

or

$$w \notin \Psi \quad \text{and} \quad z \notin \Psi$$

In either case it follows from Theorem 4.2.1 that nodes w and z will have identical entries in the row of the basis inverse corresponding to x_{1j}. Q.E.D.

The importance of the above discussion lies in its implication that a *reduced basis inverse* for RNP, which contains one row for each component of x_1^B and one column for each macro-node, is all that is required to generate the nontrivial elements of the penalty subproblem constraints. Thus the problems GP(Γ), BGP(Γ) and EOP(Γ) can be easily constructed if the following labels are obtained from the optimal basis forest for RNP.

4.2.4. Definition

For each node w in the basis forest for RNP,

$$\eta(w) = \text{the number of the macro-node to which } w \text{ belongs.}$$

For each macro-node z in the macro-node tree for RNP,

$\lambda(i, z) =$ the element of the reduced basis inverse for RNP associated with the ith component of x_1^B and the macro-node z.

4.2.5. Algorithm

Let $[\delta(n), \mu(n), \gamma(n), \alpha(n)]$ be the labels of an optimal basis forest for RNP as defined in Definition 4.1.2. Then the labels $\eta(n)$, and $\lambda(i, k)$ can be obtained as follows:

STEP 0. Set the next available macro-node $\hat{k} = 1$ and $\lambda(i, k) = 0$ for all i and k.

STEP 1. Scan sequentially the nodes until a new tree base (i.e. a node n with $\delta(n)=0$) is found. If none is found, stop; the algorithm is complete. Otherwise, set the arc index set $\Lambda = \Phi$, the current node n' = the number of the node which is the new base, $\eta(n') = 1$, and the current macro-node $k' = 1$, and go to Step 2.

STEP 2. Proceed up by letting $n = \mu(n')$. If $n = 0$ go to Step 5. Otherwise, proceed to Step 3 if $\alpha(n)$ is a component of x_1 and to Step 4 if it is a component of x_2.

STEP 3. Let $\Lambda = \Lambda \cup \{\alpha(n)\}$ if $\alpha(n)$ is oriented away from the base of the forest, and $\Lambda = \Lambda \cup \{-\alpha(n)\}$ if $\alpha(n)$ is oriented toward the base of the forest. Also let $\hat{k} = \hat{k} + 1$, $k' = \hat{k}$, $\lambda(i, k') = +1$ *for* all i in Λ such that $i > 0$, and $\lambda(-i, k') = -1$ for all i in Λ such that $i < 0$. Then go to Step 4.

STEP 4. Set $n'=n$ and $\eta(n')=k'$. Then go to Step 2.

STEP 5. Proceed right by setting $n=\gamma(n')$. If $n=0$ go to Step 6. Otherwise remove $\pm\alpha(n')$ from Λ if $\alpha(n')$ is a component of x_1, set $k'=\eta(\delta(n'))$, and then go to Step 3 if $\alpha(n)$ is a component of x_1, and to Step 4 if it is a component of x_2.

STEP 6. Proceed down by setting $n=\delta(n')$. If $n=0$, go to Step 1. Otherwise remove $\pm\alpha(n')$ from Λ if $\alpha(n')$ is a component of x_1, set $n'=n$ and $k'=\eta(n')$, and go to Step 5.

Table 1 illustrates the algorithm for the case of Figure 1.

TABLE 1. *Steps in Algorithm 4.2.5 for Example Problem RNP*

Algorithm step	Variables assigned values	Algorithm step	Variables assigned values
0	$k=1$, all $\lambda(i,k)=0$	2	$n=0$
1	$\Lambda=\Phi$, $n'=1$, $\eta(1)=1$, $k'=1$	5	$n=0$
2	$n=0$	6	$n=3$, $n'=3$, $k'=1$
5	$n=0$	5	$n=0$
6	$n=0$	6	$n=0$
1	$\Lambda=\Phi$, $n'=3$, $\eta(3)=1$, $k'=1$	1	$\Lambda=\Phi$, $n'=7$, $\eta(7)=1$, $k'=1$
2	$n=2$	2	$n=4$
4	$n'=2$, $\eta(2)=1$	3	$\Lambda=\{-17\}$, $\hat{k}=3$, $k'=3$, $\lambda(17,3)=-1$
2	$n=6$	4	$n'=4$, $\eta(4)=3$
3	$\Lambda=\{-12\}$, $\hat{k}=2$, $k'=2$, $\lambda(12,2)=-1$	2	$n=0$
4	$n'=6$, $\eta(6)=2$	5	$n=9$, $\Lambda=\Phi$, $k'=1$
2	$n=0$	3	$\Lambda=\{+11\}$, $\hat{k}=4$, $k'=4$, $\lambda(11,4)=+1$
5	$n=5$, $\Lambda=\Phi$, $k'=1$	4	$n'=9$, $\eta(9)=4$
4	$n'=5$, $\eta(5)=1$	2	$n=0$
2	$n=0$	5	$n=0$
5	$n=0$	6	$n=7$, $\Lambda=\Phi$, $n'=7$, $k'=1$
6	$n'=2$, $k'=1$	5	$n=0$
5	$n=8$, $k'=1$	6	$n=0$
4	$n'=8$, $\eta(8)=1$	1	Stop

4.3. Right-hand-sides and Objective Function

Once penalty problem constraints are generated according to the above principles, the only remaining elements of the problems to be produced are the right-hand-sides and the objective function. Review of (13), (14), and (17) will demonstrate that, like the constraint matrix, these elements of the subproblem are derived directly from the optimal basic solution to RNP. The right-hand-sides are obtained from the optimal $\overline{\text{FCNP}}$ value of y which is in turn equal to the optimal RNP flow on x_1. Similarly, if the optimal RNP simplex multiplier for node w is $\pi(w)$ and arc x_{ij} runs from node k_1 to node k_2, then the objective function coefficients are calculated by $\bar{c}_{ij}=c_{ij}+\pi(k_1)-\pi(k_2)$.

4.4. Solving Penalty Subproblems

The algorithm and computational results which follow are based on the solution of one- and two-row versions of the penalty subproblems GP(Γ), BGP(Γ), and EOP(Γ). For the one-row cases of these problems, solution is elementary. It is easy to show (see [15]) that $\nu[\text{GP}(i)]=\nu[\text{EOP}(i)]$.

Thus a solution to either of these problems can be obtained by solving two linear knapsack problems. One corresponds to equating row i of (14) to $\bar{y}_i$ (i.e. moving y_i "down" to 0), and the second corresponds to equating the row to $(u_j - \bar{y}_i)$ (i.e. moving y_i "up" to u_i). The well-known list search procedure for solving such knapsack problems reduces in the network case to finding the variable which has the smallest objective function coefficient and a $+1$ or -1 in the constraint respectively.

Only a few modifications are required in the bounded case of BGP(i). A solution to BGP(i) will continue to equal the solution to one of two linear knapsack problems like those in EOP(i). However, upper bounds on the perturbation variables are observed, so that more than one such variable may be positive in the knapsack problem solutions.

Two-row problems EOP(i, j) are solved analogously. In this case four, two-row linear programs corresponding to the right-hand-sides

$$\begin{pmatrix} \bar{y}_i \\ \bar{y}_j \end{pmatrix}, \quad \begin{pmatrix} \bar{y}_i \\ u_j - \bar{y}_j \end{pmatrix}, \quad \begin{pmatrix} u_i - \bar{y}_i \\ \bar{y}_j \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} u_i - \bar{y}_i \\ u_j - \bar{y}_j \end{pmatrix}$$

are solved. The total unimodularity of the constraint matrix for (14) can be exploited to extend the list search solution procedure outlined above to such two-row linear programs. Details are given in [14].

5. STATEMENT OF THE ALGORITHM

The computational analysis presented in the next section compares three different approaches to using penalty problems in a branch-and-bound procedure for FCNP. Letting

β (any minimization problem) = the best currently available lower bound on the value of an optimal solution to the problem,
ν (any unbounded problem) $= -\infty$
ν (any infeasible problem) $= +\infty$, and
ν (the best known solution for FCNP) $= \nu^*$,

the algorithm used in obtaining the computational results is as follows:

STEP 0. Place the whole problem FCNP in the *candidate list* (i.e. in the set of restricted versions of FCNP which might still yield an optimal solution to the full problem). Set $\beta(FCNP) = -\infty$ and $V^* = +\infty$, and proceed to Step 1.

STEP 1. Choose as the current candidate, FCNP$_c$, the element of the candidate list satisfying

$$\beta(\text{FCNP}_c) = \min\{\beta(\text{FCNP}_{c'}) : \text{FCNP}_{c'} \text{ in candidate list}\},$$

and proceed to Step 2.

STEP 2. Solve the continuous relaxation of *FCNP$_c$*, i.e. $\overline{\text{FCNP}_c}$, by solving RNP$_c$. If $\nu(\overline{\text{FCNP}_c}) \geqslant \nu^*$, proceed to Step 11 because no *completion* of FCNP$_c$ (i.e. no setting of the y_j not assigned values in FCNP$_c$) can produce a solution to FCNP with value less than that of a known solution. If $\nu(\overline{\text{FCNP}_c}) < \nu^*$, proceed to Step 3.

STEP 3. Create a feasible solution for FCNP by rounding "up" the optimal solution for $\overline{\text{FCNP}_c}$, i.e. by setting

$$s_j = \begin{cases} u_{1j} - \bar{x}_{1j} & \text{if } \bar{x}_{1j} > 0 \\ 0 & \text{otherwise} \end{cases}$$

$$y = s + \bar{x}_1$$

$$x_1 = \bar{x}_1$$

$$x_2 = \bar{x}_2,$$

where $\bar{x}_1$ and $\bar{x}_2$ are the optimal values of x_1 and x_2 in the solution of $\overline{\text{FCNP}}_c$. If the value of this rounded solution is less than ν^*, proceed to Step 4. Otherwise to go Step 5.

STEP 4. A new *incumbent solution* has been found, i.e. the rounded solution to $\overline{\text{FCNP}}_c$ provides a feasible solution to FCNP with value less than any solution found so far. Save this incumbent as a possible optimal solution, and eliminate from the candidate list any problems with β value greater than or equal to the value of the new incumbent. If the new $\nu^* = -\infty$, stop; FCNP is unbounded. Otherwise, proceed to Step 5.

STEP 5. If no group-based penalty problems are to be used, choose a branching variable y_i randomly, i.e. randomly choose a new $y_{\hat{i}}$ to fix in FCNP$_c$. Then proceed to Step 10.

If penalty problems are to be used, execute Algorithm 4.2.5 to identify the macro-node assignments $\eta_c(n)$, and the reduced basis inverse entries $\lambda_c(i, k)$ from the optimal basis forest for RNP$_c$. Then go to Step 6.

STEP 6. Construct and solve a one-row group-based penalty problem for each row i in (3–6). If the value of an optimal solution to any of these one-row problems is greater than or equal to ν^*, go to Step 11 and fathom. Otherwise, go to Step 7.

STEP 7. Using the "down" and "up" penalties obtained in the solution of the one row problems, i.e. the solutions to the knapsack problems corresponding to moving y_i "down" to 0 or "up" to u_{1i}, define

$$Y = \left\{ i : \begin{array}{l} \text{the maximum of the values of the "up" and the "down" case in the } i\text{th problem of} \\ \text{Step 6 is among the } t \text{ greatest values } (t \text{ is a predefined parameter satisfying } t \geq 1). \end{array} \right\}$$

Then go to Step 8.

STEP 8. For each $i \in Y$, select another row j_i, and construct and solve EOP$_c(i, j_i)$. If the value of an optimal solution to any of these penalty problems is greater than or equal to ν^*, go to Step 11 and fathom. Otherwise proceed to Step 9.

STEP 9. Choose the branching variable $y_{\hat{i}}$ so that $\hat{i}$ is the $i \in Y$ which maximizes

$$\max\,\{\nu[\text{EOP}_c(i, j_i) : y_i = 0],\ \nu[\text{EOP}_c(i, j_i) : y_i = u_{1i}]\}.$$

Then go to Step 10.

STEP 10. Replace FCNP$_c$ in the candidate list by two more restricted problems. One is defined by FCNP$_c$ with the additional constraint that the branching variable $y_{\hat{i}} = 0$, and the other problem is identical except that y_i is restricted to equal u_{1i}. β values for these two new candidates are as obtained from the penalty problems of Steps 6 and 8. Next proceed to Step 1 to select a new FCNP$_c$.

STEP 11. *Fathom* FCNP_c, i.e. eliminate FCNP_c from the candidate list because no completion of it can produce a feasible solution to FCNP with value less than ν^*. If the candidate list is now empty, stop; if an incumbent solution exists, it is an optimal solution for FCNP, and otherwise FCNP is infeasible. If the candidate list is not empty, proceed to Step 1 to select a new FCNP_c.

6. COMPUTATIONAL ANALYSIS

In order to learn more about the effectiveness of the algorithm of section 5, a number of randomly-generated problems were solved on Georgia Tech's Univac 1108. Answers to two general questions were of interest.

1. Would the procedure solve problems as large or larger than those previously reported in the literature?

2. Is the effectiveness of the procedure significantly changed by the use of more sophisticated penalty achemes?

6.1. Description of Experiments

The approach selected to accomplish an empirical analysis of these questions was a classical factorial experimental design. Different solution approaches were applied to randomly-generated test problems possessing all combinations of the properties previous researchers have indicated most affected computational efficiency of algorithms for fixed charge problems.

In particular, a version of the algorithm of section 5 was used to generate and solve fixed charge network problems in manners specified by the following factors:

1. *Type of problem.*—Whether the problem is a general FCNP (GNP) or a fixed charge transportation problem (FCTP). (See for example [11] for the formulation of this special case).

2. *Size of* x_1.—The number of arcs in the problem with fixed charges (code 0 = 20, code 1 = 50, code 2 = 75, code 3 = 100).

3. *Relative size of fixed costs.*—Whether the fixed costs in a problem are small or large relative to variable costs (code 1 = small, i.e. fixed charges made up less than 5 percent of the value of an optimal solution; code 2 = large, i.e. fixed charges make up 15–30 percent of the value of an optimal solution).

4. *Solution method.*—The combination of group-related penalty techniques used in solution of the problem (code 0=use no group-related techniques; code 1=use only the EOP(i); code 2=use the BGP(i) and EOP(i, j) chosen by the criteria of [15]).

The generation procedure (detailed in [14]) is an extension of the approach of Klingman, Napier and Stutz [12] which creates feasible network problems of given characteristics. Fixed charges on arcs are correlated with upper bounds.

It was initially planned to test all combinations of the above factors at the indicated level codes. However, preliminary testing revealed that structures of GNP's and FCTP's were so different that results for different solution procedures could not be compared across problem types. In addition, early results showed that problems with the dimension of x_1 greater than 50 could not be solved within reasonable time limits without some penalty techniques being used.

Thus two replications of four separate factorial experiments were actually performed. The two principle experiments focused separately on the GNP and FCTP cases. Each case was tested in all combinations of large x_1 sizes (codes 1, 2, and 3), relative fixed costs, and group-related solution methods.

In addition, two special experiments were run to analyze the impact on smaller GNP's and FCTP's of eliminating all group-related penalty schemes. These special experiments tested GNP's and FCTP's with 20 fixed charge arcs (code 0). Results for each cell in the four experiments are shown in Table 2.

TABLE 2. *Results of Experiments*

Problem type	Experiment* options	Number of nodes	Number of arcs	Fixed charge arcs	Average part solution fixed charges	Average number candidate problem solved	Average solution time (sec)
General network problems	010	18	69	20	0.021	39	1.4
	011	18	69	20	.024	5	.4
	020	18	69	20	.128	226	5.0
	021	18	69	20	.381	7	.5
	111	34	159	50	.010	6	2.2
	112	34	159	50	.010	3	1.5
	121	34	159	50	.067	10	3.2
	122	34	159	50	.198	8	4.2
	211	50	238	75	.024	13	6.8
	212	50	238	75	.031	14	16.9
	221	50	238	75	.074	12	8.3
	222	50	238	75	.076	7	7.2
	311	66	317	100	.026	18	21.6
	312	66	317	100	.029	16	22.8
	321	66	317	100	.277	38	50.8
	322	66	317	100	.098	8	11.4
Fixed charge transportation problems	010	14	33	20	.015	450	6.9
	011	14	33	20	.023	4	.2
	020	14	33	20	.202	1,542	29.6
	021	14	33	20	.140	12	4
	111	20	69	50	.020	3	.5
	112	20	69	50	.018	7	1.0
	121	20	69	50	.189	332	43.8
	122	20	69	50	.208	104	15.7
	211	24	98	75	.020	14	3.2
	212	24	98	75	.024	20	4.6
	221	24	98	75	.202	652	146.2
	222	24	98	75	.188	180	51.1
	311	26	125	100	.022	33	10.6
	312	26	125	100	.024	31	14.4
	321	26	125	100	.230	1,358	557.0
	322	26	125	100	.154	279	151.2

*In each case the first digit is the "size of $-x_1$" level, the second is the "relative size of fixed cost" level, and the third is the "solution method" level.

6.2. Analysis of Experimental Results.

Turning first to the general effectiveness of the procedure, inspection of the results in Table 2 suggest that relatively large fixed charge network problems can be solved in several minutes by either of the group-related penalty methods tested. Averages for each problem type, x_1 size and fixed charge pattern reported in Table 2 are within such reasonable computational boundaries, yet the problems with 100 fixed charge arcs are as large or larger than any FCNP's previously reported solved efficiently

In order to more precisely address the second issue of the difference between techniques, as well

as to determine the effect of various factors on the experimental results, statistical analysis of variance was applied to the results. The statistical assumptions underlying the analysis of variance could not be verified in the relatively unstructured domain of randomly-generated optimization problems. However, since all the problem factors other researchers have indicated had significant effects on computational results were included in the experiments and the ANOVA procedure is well-known to be robust, the procedure was considered adequate for indicating the importance of various effects.

The response variable selected for analysis is the number of candidate problems explicitly investigated in solving a FCNP. This variable was chosen because it appears to give the most accurate measure of the true impact of different group-related penalty procedures. Solution times are also very important, but the effect of the various penalty procedures on solution times is clouded by the programming efficiency of routines to execute the penalty procedures.

Results of the analysis of variance for this response variable are given in Tables 3 through 6 and illustrated in Figures 3 and 4. The significance of various factors implied by the results is discussed in the next several subsections.

TABLE 3. *Analysis of Variance for General Problems in Special Analysis*

Effect	Sum of squares	Degrees of freedom	Mean square	F-ratio
Relative fixed cost	17,860	1	17,860	1.13
Solution method	32,004	1	32,004	[1]2.02
Cost-method interaction	17,133	1	17,133	1.08
Error	63,308	4	15,827	1.00

[1]Significant at $\alpha = 0.25$ level.

TABLE 4. *Analysis of Variance for Transporation Problems in Special Analysis*

Effect	Sum of squares	Degrees of freedom	Mean square	F-ratio
Relative fixed cost	605,000	1	605,000	[1]55.1
Solution method	1,955,288	1	1,955,288	[1]178.2
Cost-method interaction	548,528	1	548,528	[1]53.3
Error	43,896	4	10,974	1.00

[1]Significant at $\alpha = 0.05$ level.

TABLE 5. *Analysis of Variance for General Test Problems*

Effect	Sum of squares	Degrees of freedom	Mean square	F-ratio
Size of x_1	690	2	345.0	[1]2.99
Relative fixed cost	30	1	30.0	.260
Solution method	273	1	273.0	[2]2.37
Size-cost interaction	115	2	57.5	.499
Size-method interaction	230	2	115.0	.998
Cost-method interaction	166	1	166.0	1.44
Error	1,613	14	115.2	1.00

[1]Significant at $\alpha = 0.10$ level.
[2]Significant at $\alpha = 0.25$ level.

TABLE 6. *Analysis of Variance for Transportation Test Problems*

Effect	Sum of squares	Degrees of freedom	Mean square	F-ratio
Size of x_1	407,133	2	203,566	1.10
Relative fixed cost	1,304,800	1	1,304,800	[1]7.08
Solution method	522,150	1	522,150	[2]2.83
Size-cost interaction	341,544	2	170,772	.927
Size-method interaction	195,023	2	97,512	.529
Cost-method interaction	531,633	1	531,633	[2]2.89
Error	2,579,504	14	184,250	1.00

[1]Significant at $\alpha = 0.05$ level.
[2]Significant at $\alpha = 0.25$ level.

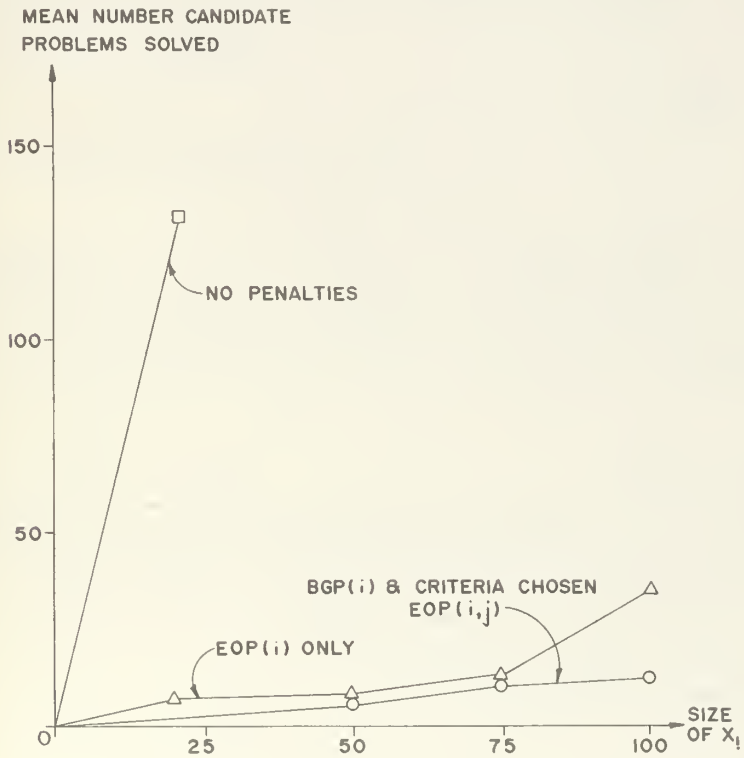

FIGURE 3. Mean number of candidate problems solved for general test problems by size of x_1 and solution method.

6.2.1. Size of x_1 Effects

Since the possible number of candidate problems increases exponentially with the size of the x_1 vector, it could be expected that the number of candidate problems actually solved would also be greatly affected by the size of x_1. Analysis of variance results for GNP's generally confirm this expectation as do the graphs in Figure 3. It is interesting, that the same statistical significance is not observed in the results for FCTP's. However, the curves in Figure 4 certainly suggest some size of x_1 effect.

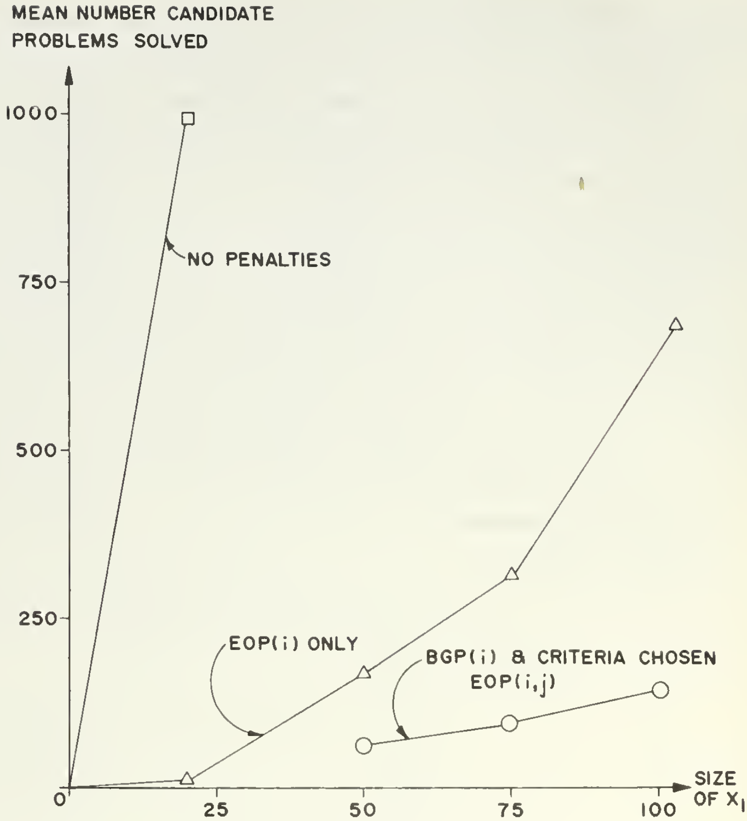

FIGURE 4. Mean number of candidate problems solved for transportation problems by size of x_1 and solution method.

6.2.2. Relative Fixed Cost Effects

Most previously reported research on fixed charge problems has indicated that computational efficiency is highly effected by the relative size of the fixed and variable costs. If fixed costs are small, $\nu(\overline{\text{FCNP}})$ provides a good estimate of $\nu(\text{FCNP})$, and only a few candidates need to be explicitly explored. When fixed costs are high, however, numerous possibilities for y must be investigated.

Experimental results for FCTP's strongly confirm this previous experience. The cost effect is very significant in both Table 4 and Table 6. However, results for GNP's show the relation between fixed and variable costs is relatively insignificant. A possible explanation of this phenomenon is the higher fixed costs in GNP's tend only to force all flows along arcs without fixed charges. Thus, the value of a $\nu(\overline{\text{FCNP}})$ as a bound on $\nu(\text{FCNP})$ is not diminished as fixed charges increase.

6.2.3. Solution Method Effects

The experimental factor of greatest interest to this research is the effect of changing the solution procedure used. Any techniques shown to be significantly superior would provide suitable focuses for future research and applications.

All results show at least mildly significant solution method effects, with the effects accentuated at high relative costs in FCTP's. The most outstanding of these effects is the difference between the no-group analysis and one-row analysis methods. Even for the relatively small case of 20 fixed charge

arcs, results in Tables 3 and 4 indicate significant improvements are obtained by using at least some group-related penalties. Moreover, review of Table 2 will show that the gain is just as great if solution time is considered instead of the number of candidates.

The difference between one-row and two-row techniques are not as clear from experimental results. Both GNP's and FCTP's showed some significant effects of variation in these solution methods (Table 5 and 6) and Figures 3 and 4 confirm the effect graphically. However, Table 2 shows cases where use of the two-row analysis increased solution time.

7. CONCLUSION

The above results are preliminary, but appear to demonstrate that group theory-based penalty approaches can be quite effectively applied to fixed charge network problems. The observations in section 4 demonstrate the significant simplifications realized in penalty approaches when the special structure of network problems is exploited. Computational experience in section 7 confirms the value of at least some use of penalties. It would certainly appear that group-based penalty approaches are worthy of significant further investigation.

REFERENCES

[1] Balinski, M. L., "Fixed Cost Transporation Problems," Nav. Res. Log. Quart., *8*, 41–54 (1961).

[2] Gomory, R. E., "On the Relation Between Integer and Non-Integer Solutions to Linear Programs," Proceedings of the National Academy of Sciences, *53*, 260–265 (1965).

[3] Gomory, R. E., "Faces of an Integer Polyhedron," Proceedings of the National Academy of Sciences, *57*, 16–18 (1967).

[4] Gomory, R. E., "Some Polyhedra Related to Combinatorial Problems," Linear Algebra and Its Applications, *2*, 451–558 (1969).

[5] Gomory, R. E. and E. L. Johnson, "Some Continuous Functions Related to Corner Polyhedra," Mathematical Programming, *3*, 23–85 (1972).

[6] Gomory, R. E. and E. L. Johnson, "Some Continuous Functions Related to Corner Polyhedra-II," Mathematical Programming, *3*, 359–389 (1972).

[7] Gray, P., "Mixed Integer Programming Algorithms for Site Selection and Other Fixed-Charge Problems Having Capacity Constraints," Technical Report No. 101, Department of Operations Research and Statistics, Stanford University, 1967.

[8] Johnson, E. L., "Programming in Networks and Graphs," ORC Report 65–1, University of California-Berkeley, 1965.

[9] Johnson, E. L., "Cyclic Groups, Cutting Planes, and Shortest Paths," presented at the Mathematical Programming Seminar, Mathematics Research Center, University of Wisconsin-Madison, 1972.

[10] Johnson, E. L. and K. Spielberg, "Inequalities in Branch and Bound Programming," IBM Research Report RC–3649, 1971.

[11] Kennington, J. L., *Fixed-Charge Transportation Problem: A Group Theoretic Approach*, dissertation in the School of Industrial and Systems Engineering, Georgia Institute of Technology, 1973.

[12] Klingman, D., A. Napier, and J. Stutz, "NETGEN: A Program for Generating Large Scale Capacitated Assignment, Transportation, and Minimum Cost Flow Problems," Management Science *20*, 814–821 (1974).

[13] Langley, R. W., *Continuous and Integer Generalized Flow Problems*, dissertation in the School of Industrial and Systems Engineering, Georgia Institute of Technology, 1973.

[14] Rardin, R. L., *Group Theoretic and Related Approaches to Fixed Charge Problems*, dissertation in the School of Industrial and Systems Engineering, Georgia Institute of Technology, 1973.

[15] Rardin, R. L. and V. E. Unger, "Some Computationally Relevant Group Theoretic Structures of Fixed Charge Problems," to appear in Mathematical Programming.

ALLOCATION OF RESOURCES TO OFFENSIVE STRATEGIC WEAPON SYSTEMS*

Kenneth D. Shere† and John W. Wingate

Naval Surface Weapons Center
White Oak Laboratory
Silver Spring, Maryland

ABSTRACT

In this paper a very versatile game model is developed for use in the long range planning of our strategic force posture. This highly aggregate model yields optimal force mixes for the triad (land- and sea-based missile systems and bombers) under a variety of constraints. The model described here is a survivability model; however, it is shown how the model can still be used as a measure of overall system effectiveness. Constraints imposed on the problem include both SALT and budget limitations.

I. INTRODUCTION

The technical review needed for the selection of strategic force postures is conducted at various levels of sophistication and complication. The most detailed and largest simulations are especially useful for the targeting of reentry vehicles and the positioning of our mobile forces. Many of the parameters in a very large scale simulation are poorly known for today's environment and must be extrapolated (to absurdity) for use in a scenario projected 5 or 10 years into the future.

Smaller scale simulations using aggregate models are useful for testing particular proposed force structures. These models require the projection of fewer parameters and are generally more believable. Generally, they are dependent on penetration tactics and postulated adversary force postures and tactics. The manager is probably unaware of all the assumptions that were made in developing the model.

An alternative approach to long range planning is the use of game theory. This approach obtains an "optimal" force posture at the expense of additional aggregation. An advantage to this approach is that the user knows all of the assumptions: however, it is subject to the criticism that not enough details are used. The results of this type of model could be used as input for a simulation model, but the primary purpose is to quickly provide the manager with both a qualitative and quantitative understanding of the effects of improving existing systems, introducing new systems, SALT and budgetary restrictions, uncontrollable increases in operating costs, etc. In the next section we develop along historical lines a game theoretic model suitable for these purposes. The model has been exercised and results are presented in the classified literature.

In section III we briefly indicate the method used to solve for the optimal allocation.

---

*This work has been supported by the Office of the Chief of Naval Operations, CNO–NOP 604, under Task Number NOL 459/R1513. Any planning ideas indicated in this paper are the authors' and do not represent Navy policy or procedures.

†Present affiliation: CTEC, Inc., 7777 Leesburg Pike, Falls Church, Va. 22043.

II. HISTORICAL DEVELOPMENT OF THE MODEL

The original "Max-Min Model" for offensive strategic systems was developed by Phipps [6]. Restricting his attention to survivability, he assumed that missile systems could be broken down into two basic types. The first class is comprised of systems that are difficult to locate but comparatively easy to destroy, once located. POLARIS is an example of such a system. Phipps coined the name "percentage vulnerable" for systems of this class because a fixed effort y_i by an attacker would bring under attack a fixed percentage of the retaliator's weapons belonging to the ith system, independent of their number. Thus, the surviving fraction of weapons (platforms) for this system is $e^{-a_i y_i}$ where a_i is the vulnerability of the system. The exponential form used here is based on Koopman's theory of random search [4].

Each of the extensions of Phipps' model, including the modifications presented here also restrict attention to survivability. The method of solution presented in the next section can be modified for any draw-down or probability of survival curve. Thus, modifications to include penetration should not involve additional computational difficulties.

The other class consists of systems which are easy to locate but difficult to destroy, e.g. MINUTEMAN. These systems were labeled "numerically vulnerable" because they find safety in numbers. Allowing fractional reentry vehicles, the fraction of weapons surviving a barrage attack is $e^{-a_i y_i/x_i}$, where x_i is the retaliator's level of effort. Temporarily we shall consider the "level of effort", x_i and y_i for each i, to be expressed in dollars.

Treating numerically vulnerable weapons as point targets, the correct survival probability curve (i.e. *not* allowing fractional reentry vehicles) is a broken line whose corners fall on an exponential curve; however, the exponential is an excellent approximation to the broken line for reasonably sized vulnerabilities a_i. Phipps [7] discusses how these vulnerability parameters can be evaluated.

For a mix of M percentage vulnerable systems and $N-M$ numerically vulnerable systems, the surviving "value" of the mix is

$$F(x, y) = \sum_{i=1}^{M} v_i x_i e^{-a_i y_i} + \sum_{i=M+1}^{N} v_i x_i e^{-a_i y_i/x_i}$$

where v_i is the value or figure of merit of the ith system—expressed originally as a cost-effectiveness parameter, e.g. megatonnage per dollar (bang per buck). Applying a conservative strategy, the objective of this model is to determine the optimal allocations x^* and y^* such that

$$F(x^*, y^*) = \min_{y} F(x^*, y) = \max_{x} \min_{y} F(x, y)$$

subject to the fiscal constraints

$$\sum_{i=1}^{N} x_i \leqslant X \qquad (x_i \geqslant 0;\ i = 1, \ldots, N)$$

and

$$\sum_{i=1}^{N} y_i \leqslant X \qquad (y_i \geqslant 0;\ i = 1, \ldots, N).$$

Unfortunately, although Phipps developed this model (just before leaving OEG–NWG) in 1960, it was not solved until Danskin [2] published his book on Max-Min in 1967. Independently, in 1965 Matheson [5] considered a Lagrange multiplier approach to solving this problem in some informal working papers and showed how the percentage vulnerable systems can be replaced by a single equivalent system. He also showed, for $N \leq 4$, how regions of the feasible set may be formulated according the mix of nonzero systems in the optimal allocation.

One of the biggest drawbacks to this model is its assumption that all systems are bought from scratch. This was all right in 1960; however, by 1965 some systems had already been procured and this should influence the results. To formulate a more realistic model, Shere and Cohen [10] took account of development or buy-in costs and Shere [8] further extended the model to also include prior investments. During the parameter evaluation and computational phases of this work, Shere noted that the cost-effectiveness parameter v_i could be adjusted to influence the results by using various methods of costing the systems.

Consequently, the model was further modified to explicitly represent future costs. Redefining x_i to be the number of retaliator's weapons (or platforms) in the ith system and y_i to be the number of search units or reentry vehicles attacking the ith system, the surviving equivalent megatonnage (EMT) of the force mix is

$$F(x, y) = \sum_{i=1}^{M} w_i x_i e^{-a_i y_i} + \sum_{i=M+1}^{N} w_i x_i e^{-a_i y_i / x_i} \tag{1}$$

where w_i is the EMT per weapon or platform. The measure of effectiveness could also be throw weight, equivalent numbers of 50 KT warheads, number of reentry vehicles, kill potential, etc. The kill potential is a measure of the effectiveness of a single reentry vehicle against targets of a specified hardness.

The attacker and retaliator need not, of course, use the same measure of effectiveness. For example, the attacker may want to allocate his resources to minimize surviving EMT, whereas the retaliator wants to maximize the surviving or residual kill potential. That is, subject to the constraints specified below we must find x^* and a function y^* such that

$$F(x, y^*(x)) = \min_{y} F(x, y) \tag{2}$$

and

$$G(x^*, y^*(x^*)) = \max_{x} G(x, y^*(x)) \tag{3}$$

where $G(x, y)$ is given by the right hand side of (1) with the EMT per weapon w_i replaced by the kill potential per weapon ϕ_i. The information available to the players requires the retaliator (x-player) to allocate his resources first, then the attacker (y-player) allocates his resources. Moreover, the retaliator knows the attacker's measure of effectiveness.

There are a variety of constraints that can be imposed. Let R_x be the total funds available to the retaliator over a specified time frame and let R_y be the funds available to the attacker. Writing the cost per system as the sum of the operating, investment and buy-in costs we have for the retaliator

$$\sum_{i=1}^{N} \left\{ o_i(x_i+\hat{x}_i)/2 + p_i \max(0, x_i-\hat{x}_i) + q_i \right\} \leq R_x \tag{4}$$

where o_i, p_i and q_i respectively denote the operating, procurement and buy-in costs per weapon in the ith system; x_i is the number of weapons at the end of the time period and $\hat{x}_i$ is the existing or initial number of weapons. Note that $q_i=0$ if either $\hat{x}_i>0$ or $x_i=0$; otherwise it is positive. As a first attempt to include the effects of phasing systems into and out of the force structure, a constant rate over the entire time period was assumed. This is reflected in the operating cost calculation. Other phase-in rates could be used. For example, operating costs of the ith system could be $o_i(x_i+3\hat{x}_i)/4$. In actual applications we have further extended this model, but are not presenting the results here, to account for (i) time lags between procurement of platforms and delivery of platforms, (ii) the IOC date for new systems, and (iii) manufacturing capabilities (both annual capacity and a minimal annual production required to keep production lines intact).

A similar constraint for the attacker is

$$\sum_{i=1}^{N} \left\{ \omega_i(y_i+\hat{y}_i)/2 + \pi_i \max(0, y_i-\hat{y}_i) \right\} \leq R_y \tag{5}$$

or, more briefly, $\Gamma(y) \leq R_y$, where $\Gamma(y)$ is the quantity (cost to the attacker) denoted by the left-hand side of (5). The attackers costs against a mobile system, for example, are the costs of operating a search and procuring more search units. Once these parameters are specified, it is not necessary to specify the particular form in which the search takes place; i.e., we can think in terms of the unit cost to an adversary to search a given area.

The attacker's development costs, in Equation (5), are ignored for several reasons. Firstly, attacker systems used in a counter-force role can generally also be used in a counter-value role (i.e., against urban/industrial targets). Consequently, the forces described in this model represent only a portion of the attacker's total forces. If a system is used in both roles, it would be impossible to specify what portion of the development cost is chargeable to one role or the other. Thus it must be realized that the attacker's budget specified, R_y, is only a portion of his total budget. Another reason that this assumption is convenient is the extreme difficulty in estimating buy-in costs. From the retaliator's viewpoint, this is a conservative assumption. Generally though, new adversary systems can be compared in terms of the vulnerability to the retaliatory system and the cost to the attacker. This is illustrated by Figure 1.

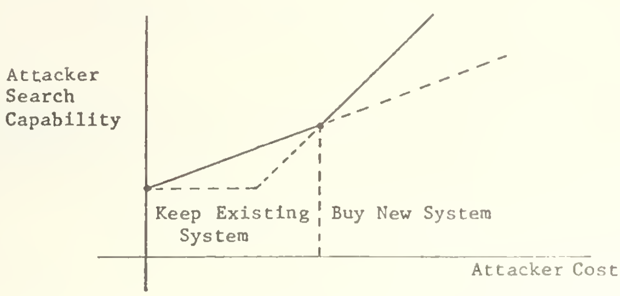

FIGURE 1. Comparison of new to existing system.

In addition to the budgetary constraints, there are arms limitations imposed by international treaties. These constraints for the retaliator on his total number of launchers, his sea-based launchers and his land-based launchers are, respectively:

$$\sum_{i=1}^{N} s_i x_i \leq S \tag{6}$$

$$\sum_{i \in M_1} s_i x_i \leq S_1 \tag{7}$$

$$\sum_{i \in M_2} s_i x_i \leq S_2 \tag{8}$$

where M_1 and M_2 are disjoint subsets of $\{1, \ldots, N\}$.

There are various other constraints also imposed on the retaliator to make the model more realistic. for example,

$$x_i = f_i x_{i-1} \qquad \text{for } i = 2, 4, 6 \tag{9}$$

where f_i is a positive constant. If x_1 were the number of Brand 1 boats at sea, then x_2 would be the number of Brand 1 boats in port. Sometimes the size of a system is influenced by political factors. Thus, there may be constraints of the type

$$x_3 \text{ is a given constant,} \tag{10}$$

$$7 \leq x_4 \leq 15 \tag{11}$$

or

$$x_6 \geq \begin{Bmatrix} 430 & \text{if } x_5 > 0 \\ 0 & \text{if } x_5 = 0 \end{Bmatrix} \tag{12}$$

Constraint (12) says that the number of weapons allocated to system 6 is at least 430 unless system 5 has been phased-out.

Hitherto only missile systems have been discussed; however, bomber systems may also be included. Dr. Alden Turner [11] of the Center of Naval Analysis recently developed an empirical formula by analyzing simulation data, for the fraction of surviving bombers of the form

$$1 - b_i \arctan a_i y_i \tag{13}$$

where y_i is the number of reentry vehicles attacking the bomber bases for bomber type i and b_i and a_i are positive numbers depending on the number of bases, reaction time and density of attack. Of

course, the fraction (13) is replaced by zero when y_i is large enough to drive the fraction negative. Using Turner's results we could replace (1) by

$$F(x, y) = \sum_{i=1}^{N_1} w_i x_i e^{-a_i y_i} + \sum_{i=N_1+1}^{N_2} w_i x_i e^{-a_i y_i / x_i} + \sum_{i=N_2+1}^{N} w_i x_i (1 - b_i \text{ arc tan } a_i y_i) \tag{1'}$$

for a complete description of the triad.

In the next section we describe the method used for solving the problem defined by (1)–(12).

III. METHOD OF SOLUTION

The *inner problem* is to find, given x, a point $y^*(x)$ which minimizes $F(x, y)$ subject to the constraints on y. The *outer problem* is to find a point x^* which maximizes $G(x, y^*(x))$ subject to the constraints on x. $G(x, y^*(x))$ is a well-defined quantity since, as will be seen later, $y^*(x)$ is unique. The existence of $y^*(x)$ and x^* is established by suitable continuity and compactness arguments.

Two approaches to optimization problems are (1) to apply analytic criteria to calculate an optimizing point, and (2) to use a numerical scheme—often an iterative one—to find a point acceptably close to the optimum. The second approach generally requires many function evaluations.

As a function of x, $G(x, y^*(x))$ is continuous and piecewise continuously differentiable. The locations of the discontinuities are not known in advance since they depend on y^*, that is, on the general solution to the inner problem. Consequently we did not find the outer problem analytically tractable. (In the special case in which only percentage vulnerable systems appear, there are no buy-in costs, and $F = G$, the function F has a saddle point (namely $(x^*, y^*(x^*))$ and a Lagrange multiplier approach can be used.) For solving the outer problem we used an iterative scheme which will be described later in this section. Quick evaluation of $G(x, y^*(x))$—and hence of $y^*(x)$—is required for such a scheme to converge in a reasonable amount of time. Lacking a general solution of the inner problem, we solved it for each x required using a Lagrange multiplier technique, obtaining an exact solution.

Solution of the Inner Problem

Let f_i be defined by:

$$f_i(x_i, y_i) = \begin{cases} w_i x_i e^{-a_i y_i} & \text{for } i = 1, \ldots, M \\ w_i x_i e^{-a_i y_i / x_i} & \text{for } i = M+1, \ldots, N \end{cases}$$

and γ_i by:

$$\gamma_i(y_i) = \frac{1}{2} \omega_i (y_i + \hat{y}_i) + \pi_i \max \{0, y_i - \hat{y}_i\}$$

so that

$$F(x, y) = \sum_{i=1}^{N} f_i(x_i, y_i)$$

and

$$\Gamma(y) = \sum_{i=1}^{N} \gamma_i(y_i) \leq R_y$$

Each f_i is convex in its second argument, and each γ_i is convex, so that the attacker is faced with a convex programming problem. Let P be the set $\{y: y_i \geq 0,\ i=1, \ldots, N\}$. Consider the problem of finding the point $y(x, \mu)$ which minimizes $F(x, y) + \mu\Gamma(y)$ over P for positive μ. Then, according to Everett's Theorem 1 [3], $y(x, \mu)$ minimizes $F(x, y)$ over P subject to the constraint:

$$\Gamma(y) \leq \Gamma(y(x, \mu)).$$

$\Gamma(y(x, \mu))$ is monotone nonincreasing in μ. From the strict convexity of $F + \mu\Gamma$ in y, it can be shown that $\Gamma(y(x, \mu))$ is continuous in μ, and that $y(x, \mu)$ is unique. The equation:

$$\Gamma(y(x, \mu)) = R_y \tag{14}$$

is solved for μ as follows:

The minimization of $F + \mu\Gamma$ with respect to y splits into N separate minimization problems, since $F + \mu\Gamma$ is separable in y, and P a cartesian product.

$$\min_{y \epsilon P} \sum_{i=1}^{N} (f_i(x_i, y_i) + \mu\gamma_i(y_i)) = \sum_{i=1}^{N} \min_{y_i \geq 0} (f_i(x_i, y_i) + \mu\gamma_i(y_i)).$$

A simple calculation then shows that

$$y_i(x_i, \mu) = \begin{cases} \min\left\{\dfrac{1}{a_i} \log \max\left\{1, \dfrac{a_i w_i x_i}{\mu \hat{s}_i}\right\}, \max\left\{\hat{y}_i, \dfrac{1}{a_i} \log \max\left\{1, \dfrac{a_i w_i x_i}{\mu s_i}\right\}\right\}\right\} & \text{for } i=1, \ldots, M \\ \min\left\{\dfrac{x_i}{a_i} \log \max\left\{1, \dfrac{a_i w_i}{\mu \hat{s}_i}\right\}, \max\left\{\hat{y}_i, \dfrac{x_i}{a_i} \log \max\left\{1, \dfrac{a_i w_i}{\mu s_i}\right\}\right\}\right\} & \text{for } i=M+1, \ldots, N \end{cases}$$

where $s_i = 1/2\ \omega_i + \pi_i$, and $\hat{s}_i = 1/2\ \omega_i$ unless $\hat{y}_i = 0$, when $\hat{s}_i = 1/2\ \omega_i + \pi_i$. See Figure 2 for a sketch of $y_i(x, \mu)$ versus $\log \mu$. This function is piecewise linear and nonincreasing in $\log \mu$, and hence $\Gamma(y(x, \mu))$ is also. The values of μ corresponding to corners of this graph are easily calculated from the condition that exactly one of the one-sided derivatives with respect to y_i of $f_i(x_i, y_i) + \mu\gamma_i(y_i)$ restricted to $[0, \infty)$ vanishes at a corner point. Equation (14) is solved by bracketing R_y between a pair of corner points and solving a linear equation. If μ^* is a solution, $y^*(x) = y(x, \mu^*)$.

Upper and lower bounds on y_i can easily be incorporated into the solution—they simply add more corner points.

As long as the solutions to the one-dimensional minimization problems can be obtained simply, as in the case presented and as in the arc tangent form used in (13), the solution of the inner problem can be reduced to the solution of one equation in one unknown. In general the equation may not be linear, so that the interpolation procedure may have to be replaced by some other method, not necessarily by one giving an exact solution.

The Outer Problem

We solved the outer problem by a two-stage process. In the first stage several hundred feasible points $x^{(1)}, \ldots, x^{(k)}$ were generated at random and $G(x^{(1)}, y^*(x^{(1)})), \ldots, G(x^{(k)}, y^*(x^{(k)}))$ evaluated.

The $N+1$ best points were used as an initial complex of points for a variation of Box's search method [1]. Both procedures are described in detail in a separate document [9].

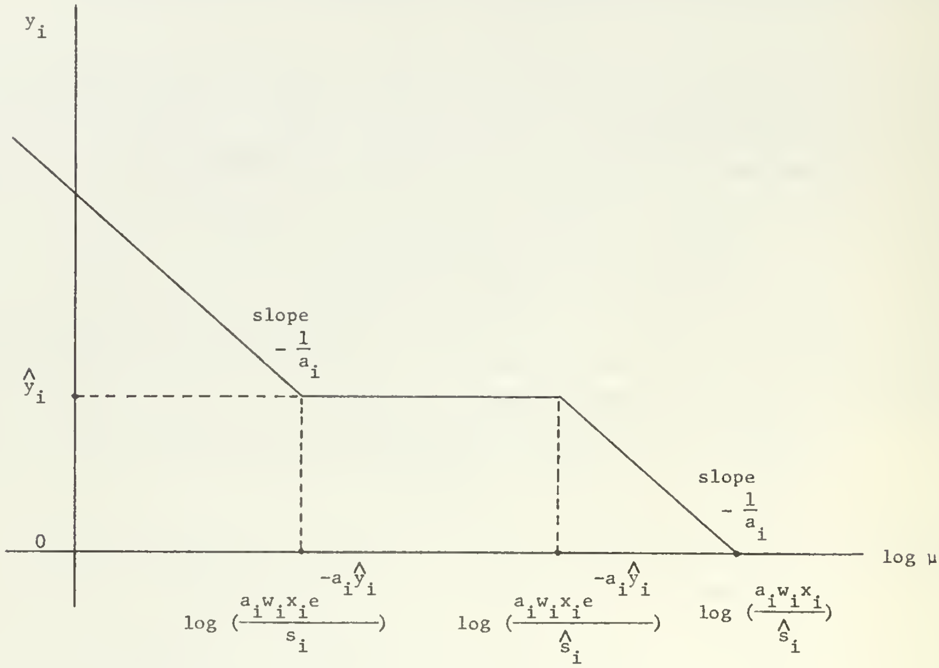

FIGURE 2. Plot of $y_i(x, \mu)$ vs. $\log \mu$ for a percentage vulnerable system.

IV. CONCLUSION

We have presented a model for allocation of resources to offensive strategic systems that can be used to realistically reflect the long range planning process.

Suppose, for example, that the planner (i.e., the retaliator) wants first to obtain a specified capability against hard targets and secondly to maximize his surviving EMT. He can use this model, by changing the objective function, to minimize cost constrained by a lower bound on surviving kill potential. Using the output from this problem as the input of the retaliator system lower bounds, he can then maximize surviving EMT, provided the minimum cost resulting from the first step is less than his budget.

The planner may have the missile for each type of system categorized as useful against hard targets or useful against soft targets. He wants to maximize surviving kill potential for the first group of missiles and to maximize surviving EMT for the second group. Being presented with a multicriterion problem, the planner must decide the type of optimum he desires. For example, he can take a convex combination of the two criteria as his objective function. Another approach is to find the Pareto optimum points. This can be done by (i) solving the "kill potential" problem ignoring surviving EMT and (ii) then solving repeatedly the "EMT" problem while imposing lower bounds on the kill potential ranging from zero to the answer obtained in step (i). The planner now has a range of Pareto optimal outcomes. He can then choose a particular force mix to advocate based on other considerations.

V. ACKNOWLEDGEMENT

The authors thank John P. Coyle (CNO–NOP 60SA) for many discussions relating to this model.

REFERENCES

[1] Box, M. J., "A New Method of Constrained Optimization and a Comparison to Other Methods," Computer Journal *8*, 42–52 (1965).

[2] Danskin, J. M., *The Theory of Max-Min* (Springer-Verlag Inc., New York, 1967).

[3] Everett, H., III, "Generalized Lagrange Multiplier Method for Solving Problems of Optimum Allocation of Resources," Operations Research *11*, 399–417 (1963).

[4] Koopman, B. O., "The Theory of Search I," Operations Research *4*, 324–346 (1956).

[5] Matheson, J. D., private communication, 1965.

[6] Phipps, T. E., "Optimum Allocation of Effort for Deterrence," Operations Evaluation Group—Naval Warfare Analysis Group (now a division of the Center for Naval Analyses, Arlington, Va. 22209), OEG–NWG 43–60, 4 May 1960.

[7] Phipps, T. E., "Optimum Systems Choice for Strategic Retaliation: An Application of Max-Min Theory," Naval Ordnance Laboratory Technical Report NOLTR 67–59, 1967.

[8] Shere, K. D., "A Resource Allocation Model with Development Costs and Prior Investments," Naval Ordnance Laboratory Technical Report NOLTR 71–189, 1971.

[9] Shere, K. D., "The Box Optimization Method," Naval Ordnance Laboratory Technical Report NOLTR 74–167, 1974.

[10] Shere, K. D. and Cohen, E. A., "A Defense Allocation Problem with Development Costs," Nav. Res. Log. Quart., *19*, 525–537 (1972).

[11] Turner, A., private communication, 1973.

PROJECT SCHEDULING: THE EFFECTS OF PROBLEM STRUCTURE ON HEURISTIC PERFORMANCE

James H. Patterson

Department of Management Science and Organizational Behavior
The Pennsylvania State University

ABSTRACT

Individual characteristics of multiple-constrained resource, project scheduling problems are examined in an attempt to predict the solution obtainable with heuristic methods. Difficulties encountered in performing this type of research are described, and several multiple regression models are developed for predicting heuristic performance. Both single and multiple project data are examined, and results reported demonstrate the efficacy of determining beforehand the method used for problem solution.

I. INTRODUCTION

Previous studies have been made of the effects of time and resource characteristics on different priority dispatch sequencing rules for resolving the conflicts that develop between the resources required by a group of activities and those available [4, 7, 11, 12]. How satisfactory the results of these studies were when analyzed statistically seems to have depended largely on whether artificial projects were generated on a computer and then extended to actual projects, or whether data obtained from actual practice were used.

When artificial projects were examined, the results were generally not very satisfactory. Davies, for example [7], made an extensive evaluation of the effects of network and resource characteristics on alternate methods of scheduling single-resource, single-project networks. Her experimental medium consisted of a set of artificially generated networks that satisfied various measures of network configuration and resource usage. Four network parameters (percentage of dummy activities, complexity, density, and resource limitation) were each varied over three levels in constructing the networks. Each network was processed in forward and in reverse order, with and without activity splitting. Two criterion functions (proportional increase in project duration and proportional usage of resources) were investigated. For both objective functions, the order of importance in the analysis of variance was: resource limitation, density, complexity, percentage of dummy activities, and the sequencing criteria (with the latter being of relatively minor importance). Unfortunately, the values of the objective functions did not vary greatly with the criteria examined. In fact, Davies reported that there was no significant difference at the 0.05 level between any pair of means for the time-based objective function.

When project data have been obtained from actual practice, the results have been encouraging [12]. There have even been interesting results when only part of the data used was obtained from practice. Davis, for example [4], isolated several summary measures of network configuration and

resource usage in an attempt to predict the percent increase in critical path duration that resulted when a minimum-late-finish-time heuristic was applied to constrained-resource problems. His experimental medium consisted of 202 different single-project networks solved under a variety of resource limitations which yielded 721 test problems. For the one sequencing rule, Davis derived several multiple regression models for predicting the percent increase in critical path duration due to limited resource availabilities. Each model had R^2 values in the neighborhood of 0.85–0.95. The standard error of estimate in these models was equally good. It ranged from about 0.07 to 0.11. These findings demonstrate that it is possible to statistically isolate project summary measures that can be used as a guide in predicting heuristic performance.

In this paper, difficulties which are often encountered when using artificial projects are described. Two groups of dissimilar project types, both artificially generated, are then investigated. Multiple (stepdown) regression is used to predict heuristic performance. Guidelines are then developed for scheduling project activity. Advantages of analyzing problem structure before choosing a technique for solving it are also reported.

II. SEQUENCING RULES EXAMINED

The priority dispatch scheduling rules examined are indicated in Table 1. These rules have all been tested previously and represent a collection of those which have been found effective elsewhere, as well as some which have generally produced poor results on selected problems. A rule which resolves resource conflicts on a purely random basis has also been included to be used as a comparison.

TABLE 1. *Scheduling Rules Examined*

Scheduling rule	Identification
Least Total Float[1,2]	LTF
Greatest Resource Demand[3]	GRD
Greatest Remaining Resource Demand[4]	GRRD
Resource Scheduling Method[5]	RSM
Shortest Imminent Operation	SIO
Greatest Resource Usage	GRU
Earliest Late Finish Time[1]	LFT
Most Jobs Possible	MJP
Random Activity Selection	RAN

[1] Determined by conventional critical path methods.
[2] The actual rule used is a dynamic version of the Least Total Float heuristic, and is equivalent to an Earliest Late Start Time rule (see [6] for a proof of this relationship).
[3] This rule is the same as the Greatest Total Resource Demand rule of [12].
[4] Used only on multiproject problems.
[5] Used only on single-project problems.

Each of the rules listed in Table 1 is applied in conjunction with the parallel method of scheduling in which sequencing priorities are determined during scheduling, rather than before. With the exception of the RSM heuristic (a description of which can be found in [1]), the heuristic title indicates the priority given to competing activities. For example, the *Least* Total Float heuristic gives highest priority to activities possessing the *least* amount of activity total float; the *Greatest* Resource Demand rule gives highest priority to those activities demanding the *greatest* amount of resources. (A more complete

description of each of the rules in table 1 can be found in [6] or in [12].) Ties in activity priority arc generally broken first by project number (for the multiproject case) and then be activity (job) number. No "add-on" or reschedule rules of the type investigated by Wiest [15–17] are included in this investigation because we assume that each activity is completed within its specified duration with a constant usage of resources. Activities cannot be expedited (slowed) with the addition (subtraction) of resource units from those specified.

III. HEURISTIC DISCRETION AND THE USE OF ARTIFICIAL PROJECTS

In order to exercise purported logic relative to a specific criterion, a heuristic scheduling rule must be able to discriminate among activities. This often leads to difficulties when data are artificially generated. Discretion by itself, however, is often not sufficient. As is shown below, the rule must have the opportunity to make resolutions of sequencing conflicts which will have an ultimate bearing on the results obtained.

Figure 1, for example, represents two project networks requiring fixed amounts of one resource. The activities in Network 1 of Figure 1 have identical values of total float, have the same duration (three time units), and require the same amounts of resources (five units for each of three periods, for 15 unit-periods). With a limit of five on the number of resource units available, application of each of the sequencing rules given in Table 1 (except on occasion, RAN) results in an identical schedule if ties are broken by lowest job number. The priority dispatch sequencing rule used to solve this problem is relatively unimportant.

It is not difficult to explain why any selection of scheduling rules results in identical schedules for the above problem. Resources are "tight" in the sense that each activity requires five units of the particular resource over its three period duration, and only five units of the resource are available. Only one activity can be on-going at any one time. Resource utilization over the constrained-resource duration is 100 percent.

The scheduling abilities of the priority dispatch rules are not a function of resource utilization alone, however. This is evident in scheduling network 1 with a resource limit of nine units. The ratio of usage to availability of this resource measured over the resource-constrained project duration (schedule span) is only 56 percent; yet the sequencing rules given in Table 1 still do not discriminate among the activities in sequencing them, and schedules identical to those obtained with a resource limit of five units prevail.

Network 2 of Figure 1 was constructed to illustrate yet another type of problem for the sequencing rules of Table 1. The majority of activities in Network 2 have different durations, demands for resources, amounts of total float, and so on. Hence, there are ample opportunities in scheduling this second project to discriminate among activities in making sequencing decisions. With a resource limit of six units, for example, application of each of the sequencing rules given in Table 1 results in a nonidentical schedule. These results are summarized in Table 2. Note also from Table 2 that use of each of the sequencing rules results in a schedule span of 31 time units, even though no two schedules are alike! An evaluation of these rules on Percent Increase in Critical Path Duration or on Project Makespan (or on a host of other criteria) would show no one rule superior or inferior to the rest.

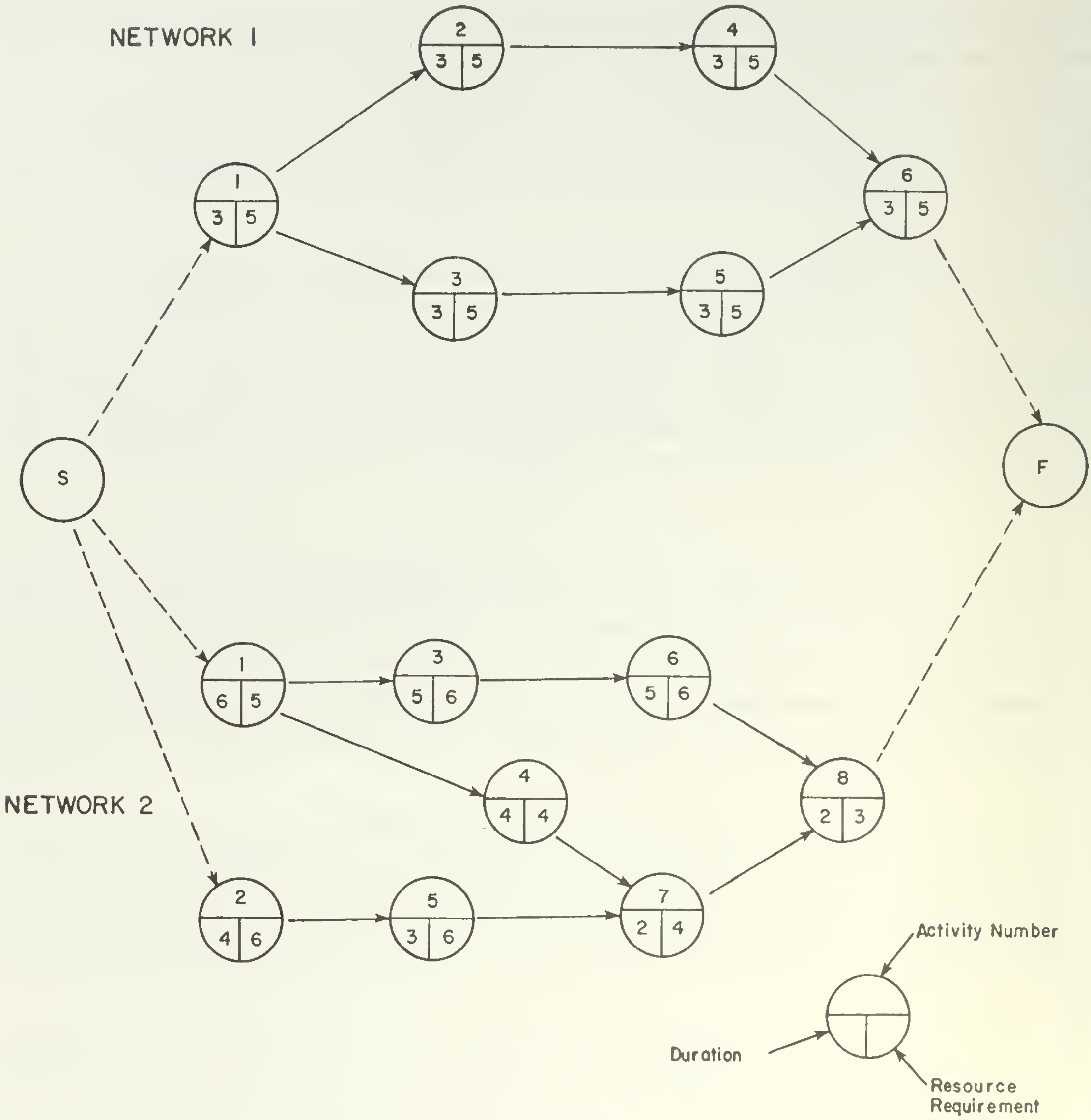

FIGURE 1. Example of project network.

TABLE 2. *Period in Which Activity Is Scheduled to Start Network 2, Limit of Six Resource Units*

Activity	Heuristic+						
	LTF	GRD	SIO	GRU	LFT	MJP	RAN
1	1	1	8	8	1	1	1
2	12	17	1	1	7	21	16
3	7	7	20	14	11	11	11
4	16	24	14	24	16	7	7
5	20	21	5	5	20	25	25
6	23	12	25	19	23	16	20
7	28	28	18	28	28	28	28
8	30	30	30	30	30	30	30

+Descriptions of each of these heuristics can be found in Table 1.

It is indeed likely that, for a host of real and artificial problems, the choice of a scheduling rule makes little difference in the results of scheduling effort. It is, therefore, relatively unimportant which sequencing rule is employed. Although the networks of Figure 1 were fabricated to produce the results shown, the results reported are *not* atypical of those obtained in examining smaller, laboratory type projects. For example, difficulties often arise when attempting to control certain values of project and resource parameters when smaller, laboratory type projects are examined. Resource utilization is one characteristic which is particularly difficult to control because of integer restrictions on the availability and use of resources. Other characteristics pose similar problems when the data are artificially generated. Considerable care must, therefore, be exercised if the results of the experiment are to be generalized to the types of projects commonly found in practice.

Project and resource summary measures designed to predict heuristic performance as well as to indicate characteristics of problems for which heuristic choice is likely to be unimportant are described in the following section. These measures are intended for use in scheduling actual as well as artificial projects and project sets.

IV. IDENTIFICATION OF INDEPENDENT VARIABLES USED IN PREDICTING HEURISTIC PERFORMANCE

In this section, the independent variables which *may* contribute to good (or poor) heuristic performance are identified. They are divided into three categories. In one, time and network based parameters are computed *prior* to critical path analysis. In the second, time and network based parameters are computed *subsequent* to critical path analysis. The third category includes resource-based parameters which are generally computed subsequent to critical path analysis. Note from Figure 1 that a multiproject scheduling problem can be treated in the same manner as a single-project scheduling problem if a dummy activity is used to precede (succeed) the beginning (ending) activities of all projects. It is useful, however, to identify several sets that simplify the notation used when single vs. multiproject parameters are intended. Table 3 gives these sets, and other notation for identifying variables used in describing the various problem characteristics examined.*

1. Time and Network Based Parameters Computed Prior to Critical Path Analysis

NPROJ	Number of Projects To Be Scheduled
NNODE	Number of Nodes (Activities) To Be Scheduled
NARC	Number of Arcs (Precedence Relationships)
NDUMMY	Number of Dummy Activities (0-duration)

*In the definitions that follow, an attempt has been made to identify parameters that have been initially described by other researchers of the constrained-resource, project scheduling problem. These are indicated by the author's name following the description of the parameter referenced. In some instances, the parameter as used in this paper may not be as originally described by the author.

TABLE 3. *Identification of Variables and Sets Used In Computation of Project Parameters*

[Independent Variables]

Notation	Description
N	The set of all activities to be scheduled.
P	The set of all projects to be scheduled.
R	The set of all resource categories.
d_{ij}	Duration of activity j of project i.
r_{ijk}	Per-period requirement of resource k by activity j of project i.
R_k	Availability of resource k in each period of the schedule span.
CP_i	Critical path length of project i.
$FF_{ij}^{(1)}$	Free-float of activity j of project i.
$TF_{ij}^{(1)}$	Total-float of activity i of project j.

[1]Determined by conventional critical path methods.

ΣDUR Sum of the Activity Durations

$$\sum_N d_{ij}$$

$\overline{\text{X}}$DUR Average Activity Duration

$$\frac{\sum \text{DUR}}{\text{NNODE}}$$

VA–DUR Variance in Activity Duration

$$\frac{\sum_N (d_{ij} - \overline{\text{X}}\text{DUR})^2}{\text{NNODE} - 1}$$

T–DENSITY Total Activity Density (Johnson)

$$\sum_N \max\{0, \text{Number of Predecessor Activities} - \text{Number of Successor Activities}\}$$

$\overline{\text{X}}$DENSITY Average Activity Density

$$\frac{\text{T–DENSITY}}{\text{NNODE}}$$

COMPLEXITY Project Complexity (Pascoe)

$$\frac{\text{NARC}}{\text{NNODE}}$$

Of those parameters identified above, Variance in Activity Duration (VA–DUR) is likely to have an effect on the performance of the SIO heuristic, which is an optimal rule (under a set of restrictive conditions) for the one-machine sequencing problem of the job shop. This parameter should also affect the performance of other rules which are based in part upon activity duration.

The last three parameters (T–DENSITY, $\overline{\text{X}}$DENSITY, and COMPLEXITY) measure the interconnectedness of a network, and thereby influence when (in terms of network logic) an activity can be scheduled.

2. Time and Network Based Parameters Computed Subsequent to Critical Path Analysis

ΣCPL Sum of the Critical Path Lengths

$$\sum_{P} CP_i$$

$\overline{\text{X}}$CPL Average Critical Path Length

$$\frac{\Sigma\text{CPL}}{\text{NPROJ}}$$

VA–CPL Variance in Critical Path Lengths

$$\frac{\sum_{P} (CP_i - \overline{\text{X}}\text{CPL})^2}{\text{NPROJ} - 1}$$

MAXCPL Maximum Critical Path Length

$$\max_{P} \{CP_i\}$$

ΣSLACK Total Slack (Float) of All Activities

$$\sum_{N} TF_{ij}$$

NSLACK Number of Activities Possessing Positive (NonZero) Total Float

$$\sum_{N} \begin{Bmatrix} 1 & \text{if } TF_{ij} > 0 \\ 0 & \text{if } TF_{ij} = 0 \end{Bmatrix}$$

PCTSLACK Percent of Activities Possessing Positive Total Slack

$$\frac{\text{NSLACK}}{\text{NNODE}}$$

$\overline{\text{XSLACK}}$ Average Total Slack Per Activity

$$\frac{\Sigma\text{SLACK}}{\text{NNODE}}$$

TOTSLACK–R Total Slack Ratio

$$\frac{\Sigma\text{SLACK}}{\text{MAXCPL}}$$

$\overline{\text{XSLACK}}$–R Average Slack Ratio

$$\frac{\overline{\text{XSLACK}}}{\text{MAXCPL}}$$

PDENSITY–T Project Density—Total

$$\frac{\Sigma\text{DUR}}{\Sigma\text{DUR} + \Sigma\text{SLACK}}$$

ΣFREESLK Free Slack (Float) of All Activities (Johnson)

$$\sum_N FF_{ij}$$

NFREESLK Number of Activities Possessing Positive (Non-Zero) Free Slack

$$\sum_N \left\{ \begin{array}{ll} 1 & \text{if } FF_{ij} > 0 \\ 0 & \text{if } FF_{ij} = 0 \end{array} \right\}$$

PCTFREESLK Percent of Activities Possessing Positive Free Slack

$$\frac{\text{NFREESLK}}{\text{NNODE}}$$

$\overline{\text{XFREESLK}}$ Average Free Slack Per Activity

$$\frac{\Sigma\text{FREESLK}}{\text{NNODE}}$$

PDENSITY–F Project Density—Free (Pascoe)

$$\frac{\Sigma\text{DUR}}{\Sigma\text{DUR} + \Sigma\text{FREESLK}}$$

Of the parameters listed in this second category, those that reflect measures of float or slack present in activities are likely to account for a significant portion of the variation present in the performance of the Least Total Float heuristic. Since measures of slack do, however, reflect scheduling freedom in the sense that specific activities can be delayed without delaying the completion of a project, this measure will undoubtedly account for a large portion of the variation in the behavior of the other sequencing rules. Delays should (on the average) be less when using (e.g.) the SIO heuristic in scheduling a project with large amounts of slack on a high proportion of activities than in scheduling a project in which few activities possess relatively small amounts of slack or total float.

This second category of measures includes parameters based on the total float *and* the free float present in project networks. Measures of free float were included because the measures of total float overstate the amount of scheduling freedom available in an activity; activity total slack may be duplicated for all activities in a given chain. For such activities, the delay in a preceding activity means a loss of slack in the succeeding activities. The above measures also reflect the percent of activities which possess either total or free slack, as well as the amounts possessed.

3. Resource Based Parameters Generally Computed Subsequent to Critical Path Analysis

PCTR_k Percent of Activities Requiring Positive Amounts of Resource k

$$\frac{\sum_N \begin{Bmatrix} 1 & \text{if } r_{ijk} > 0 \\ 0 & \text{if } r_{ijk} = 0 \end{Bmatrix}}{\mathrm{NNODE}} \quad \text{for all } k \epsilon R$$

MIN%DEMAND Minimum Percent of Demands for a Resource

$$\min_R \{\mathrm{PCTR}_k\}$$

$\overline{\mathrm{X}}$%DEMAND Average Percent of Demands for Resources

$$\frac{\sum_R \mathrm{PCTR}_k}{\mathrm{NRES}}$$

MAX%DEMAND Maximum Percent of Demands for a Resource

$$\max_R \{\mathrm{PCTR}_k\}$$

UTIL_k Utilization of Resource k (Measured over the longest critical path length) (Davis)

$$\frac{\sum_N r_{ijk} \cdot d_{ij}}{R_k \cdot MAXCPL} \quad \text{for all } k \epsilon R$$

MINUTIL Minimum Resource Utilization

$$\min_{R} \{\mathrm{UTIL}_k\}$$

$\overline{\mathrm{X}}$UTIL Average Resource Utilization

$$\frac{\sum_{R} UTIL_k}{\mathrm{NRES}}$$

MAXUTIL Maximum Resource Utilization

$$\max_{R} \{\mathrm{UTIL}_k\}$$

DMND_k Average Quantity of Resource k Demanded When Required By An Activity

$$\frac{\sum_{N} r_{ijk}}{\sum_{N} \begin{Bmatrix} 1 \text{ if } r_{ijk} > 0 \\ 0 \text{ if } r_{ijk} = 0 \end{Bmatrix}} \text{ for all } k \epsilon R$$

$\bar{\mathrm{X}}$DMND Average Quantity of Resources Demanded When Demanded

$$\frac{\sum_{R} \mathrm{DMND}_k}{\mathrm{NRES}}$$

CONSTR_k Resource Constrainedness

$$\frac{\mathrm{DMND}_k}{R_k} \quad \text{for all } k \epsilon R$$

MINCON Minimum Resource Constrainedness

$$\min_{R} \{\mathrm{CONSTR}_k\}$$

$\bar{\mathrm{X}}$CON Average Resource Constrainedness

$$\frac{\sum_{R} \mathrm{CONSTR}_k}{\mathrm{NRES}}$$

MAXCON Maximum Resource Constrainedness

$$\max_{R} \{\mathrm{CONSTR}_k\}$$

VA–CON Variance in Resource Constrainedness

$$\frac{\sum_{R} (\text{CONSTR}_k - \bar{\text{X}}\text{CON})^2}{\text{NRES} - 1}$$

TCON_k Resource Constrainedness Over Time

$$\frac{\sum_{N} r_{ijk} \cdot d_{ij}}{\left[\sum_{N} \begin{cases} 1 & \text{if } r_{ijk} > 0 \\ 0 & \text{if } r_{ijk} = 0 \end{cases}\right] \cdot \left[R_k \cdot \text{MAXCPL} \right]} \quad \text{for all } k \epsilon R$$

MINCON–TM Minimum Resource Constrainedness Over Time

$$\min_{R} \{\text{TCON}_k\}$$

X̄CON–TM Average Resource Constrainedness Over Time

$$\frac{\sum_{R} \text{TCON}_k}{\text{NRES}}$$

MAXCON–TM Maximum Resource Constrainedness Over Time

$$\max_{R} \{\text{TCON}_k\}$$

VA–CON–TM Variance In Resource Constrainedness Over Time

$$\frac{\sum_{R} (\text{TCON}_k - \bar{\text{X}}\text{CON–TM})^2}{\text{NRES} - 1}$$

ACON_k Resource Constrainedness Using All Activities as a Base

$$\frac{\sum_{N} r_{ijk}}{\text{NNODE} \cdot \text{R}_k} \quad \text{for all } k \epsilon R$$

MINCON–ALL Minimum Resource Constrainedness Using All Activities as a Base

$$\min_{R} \{\text{ACON}_k\}$$

$\bar{\text{X}}$CON–ALL Average Resource Constrainedness Using All Activities as a Base

$$\frac{\sum_{R} \text{ACON}_k}{\text{NRES}}$$

MAXCON–ALL Maximum Resource Constrainedness Using All Activities as a Base

$$\max_{R} \{\text{ACON}_k\}$$

VA–CON–ALL Variance in Resource Constrainedness Using All Activities as a Base

$$\frac{\sum_{R} (\text{ACON}_k - \bar{\text{X}}\text{CON–ALL})^2}{\text{NRES}-1}$$

The above resource utilization parameters reflect the "tightness" of certain resource types. Obviously, if the demand for a particular resource at any point in time does not exceed the availability, R_k, then this resource is not very constraining and few resolutions of conflicts in the demand for this resource will have to be made. And where conflicts do have to be resolved, but the quantities required approach the availability of a resource, conflict resolution will have to be made but will be of little consequence in terms of the ultimate duration of a project. An example of this latter situation was given in the previous section.

Between the extremes of a large portion of the activities demanding a large quantity of the availability of a resource and resources being available to schedule all competing activities without resolution, a given heuristic has the potential to effect decisions which may bear heavily on the criterion being evaluated. This is because, of course, the heuristic has the ability to select some (possibly unique) subset of the activities available for scheduling. As the number of activities which could be included in a given subset of scheduled activities increases, and as the number of feasible subsets of activities for selection increases, the more potential there is for effecting decisions which will have an impact on the final results. The resource parameters herein termed "Constrainedness" and "Constrainedness Over Time" are examples of quantifiable indices of potential decision-making effectiveness.* High values of certain constrainedness parameters imply the potential for masking the intended effectiveness of heuristic procedures. As the average demand for a particular resource decreases, for example, the potential for making effective decisions increases. Several different constrainedness indices are included in order to identify their potential effect on heuristic performance.

OFACT_k Obstruction Factor of Resource k. (Davis, Pascoe)

$$\frac{\text{"Excess" Resource Requirement}_k}{\text{Resource Work Content}_k} \quad \text{for all } k \in R$$

*The importance of the constrainedness parameters in job shop scheduling research should be apparent. The implication of a job (activity) using all of an available resource (man, machine, etc.) is, of course, that the resolution of all resource conflicts implies *one* job or activity is scheduled and *all* other available jobs must be postponed: no two (or more) activities using the same resource can be on-going at the same time.

$$\text{Excess Resource Requirement}_k = \sum_{\text{MAXCPL}} \max \{0, \text{Demand}_k - R_k\}$$

where the demand for resource k is based on an all early start schedule. (See Figure 2.)

$$\text{Resource Work Content}_k = \sum_{N} r_{ijk} d_{ij}$$

TOTOFACT Total Obstruction Factor (Davis)

$$\sum_{R} \text{OFACT}_k$$

MINOFACT Minimum Obstruction Factor

$$\min_{R} \{\text{OFACT}_k\}$$

MAXOFACT Maximum Obstruction Factor

$$\max_{R} \{\text{OFACT}_k\}$$

UFACT_k Underutilization Factor

$$\frac{\text{Underutilization}_k}{\text{Total Work Content}_k} \quad \text{for all } k \in R$$

$$\text{Underutilization}_k = \sum_{\text{MAXCPL}} \max \{0, R_k - \text{Demand}_k\}$$

where the demand for resource k is based on an all early start schedule (see Figure 2).

TOTUFACT Total Underutilization Factor

$$\sum_{R} \text{UFACT}_k$$

MINUFACT Minimum Underutilization Factor

$$\min_{R} \{\text{UFACT}_k\}$$

MAXUFACT Maximum Underutilization Factor

$$\max_{R} \{\text{UFACT}_k\}$$

NOVER_k Number of Time Periods The Demand for Resource k Exceeds the availability of Resource k (where the demand is based on an all early start schedule)

$$\sum_{\mathrm{MAXCPL}} \begin{cases} 1 & \text{if Demand}_k > R_k \\ 0 & \text{if Demand}_k \leq R_k \end{cases} \quad \text{for all } k \in R$$

$\bar{\mathrm{X}}\mathrm{OVER}$ Average Excess Demand Time Periods for Resources

$$\frac{\sum_R \mathrm{NOVER}_k}{\mathrm{NRES}}$$

MINOVER Minimum Excess Demand Time Periods for Resources

$$\min_R \{\mathrm{NOVER}_k\}$$

MAXOVER Maximum Excess Demand Time Periods for Resources

$$\max_R \{\mathrm{NOVER}_k\}$$

NUNDER_k Number of Time Periods the Availability of Resource k Exceeds or Equals the Demand for Resource k (where the demand is based on an all early start schedule)

$$\sum_{\mathrm{MAXCPL}} \begin{cases} 1 & \text{if } R_k \geq \text{Demand}_k \\ 0 & \text{if } R_k < \text{Demand}_k \end{cases} \quad \text{for all } k \in R$$

$\bar{\mathrm{X}}\mathrm{UNDER}$ Average Time Underutilization of Resources

$$\frac{\sum_R \mathrm{NUNDER}_k}{\mathrm{NRES}}$$

MINUNDER Minimum Time Underutilization of Resources

$$\min_R \{\mathrm{NUNDER}_k\}$$

MAXUNDER Maximum Time Underutilization of Resources

$$\max_R \{\mathrm{NUNDER}_k\}$$

Figure 2 is a resource profile of the demands for the one resource involved in Network 2 of Figure 1 based upon each activity being scheduled at its critical path analysis determined early start time. As

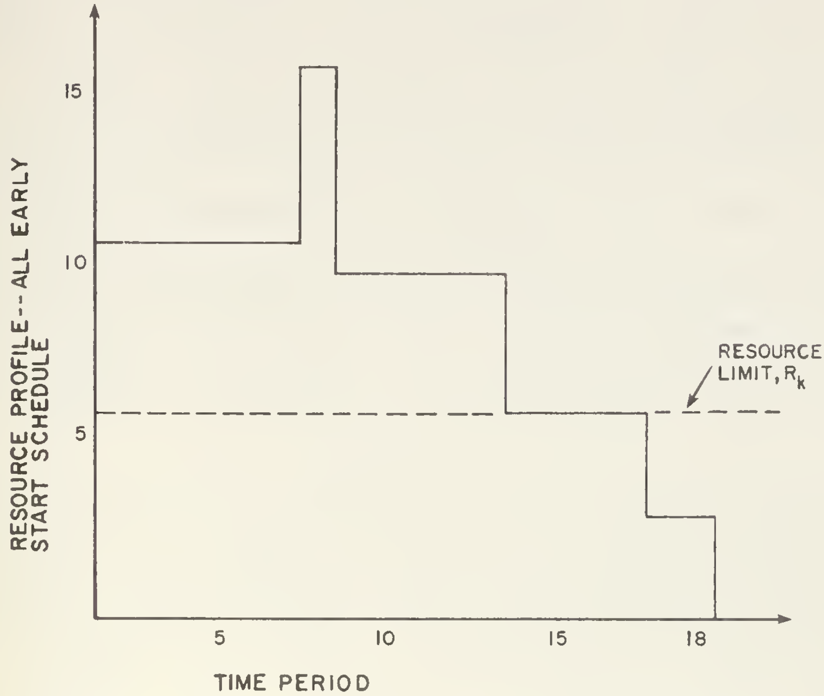

FIGURE 2. Resource profile of an all early start schedule for network 2 of figure 1.

shown in the nonresource constrained version of this problem, the peak demand for resources occurs in time period seven at 16 units and reaches a low demand of three units in time periods 17 and 18. An imposed limit of six units on the quantity of resource available will extend the duration of this project beyond the 18 time periods indicated by its critical path length. These latter resource based parameters provide an indication of the conflicts which will develop because of the limitation on resources. These measures assess both the number of time periods in which resources are underutilized or overutilized, and the amounts of overutilization and underutilization based on an all early start schedule. Knowing in advance, for example, that there are very few time periods in which the demand for resources exceeds the availability, one might be tempted to employ conventional critical path procedures and resolve the conflicts as they develop in the life of the project and not plan the sequencing of activities with any formal heuristic procedure.

A FORTRAN program was written to calculate the parameters in each of the three categories above. The values obtained then served as independent variables in a regression model to predict heuristic performance. Stepdown multiple regression was used to analyze the scheduling results, and independent variables with a net regression coefficient significant (t-test) at the 95 percent level remained in the regression equations developed.

V. MULTIPLE REGRESSION RESULTS FOR PREDICTING HEURISTIC PERFORMANCE

Sixty multiproject scheduling problems were computer generated using network, time, and resource parameters from [12] to construct each network. Project sets generated consist of 6 to 10

projects each, and each project consists of 20 to 40 activities. Thirteen different resource categories were involved, and each activity demanded fixed, positive amounts of resources from as many as 13 resource categories. This generated data is thus representative of that found in practice.

Three criterion functions are investigated for this multiproject data: (1) minimize the sum of the delays beyond the critical path length for all projects; (2) minimize the sum of the total weighted delays of the projects, where the weights are determined by the size of the project measured by total resource-unit requirements (total work content) for project completion; and (3) minimize the percent increase in critical path duration, where the group of projects are conjoined by dummy nodes to form one project. This latter criterion is equivalent to minimizing makespan, or the time required to complete all jobs (projects). In a typical multiproject organization, the firm rarely accepts a group of projects, finishes this group, then accepts another group, finishes it, etc. Rather, projects are entering and being completed in the organization coterminately; still others are in progress. For completeness, however, and to contrast our results in minimizing project makespan with those results reported in job shop scheduling research and elsewhere, the criteria of minimizing project makespan is included in the analysis.

Table 4 presents some summary statistics for the 60 project sets generated, and Table 5 gives information on a factor analysis performed to combine the project variables into independent factors. The six factors retained account for 71 percent of the information accounted for by all independent variables. Tables 6, 7, and 8 then give the results of the regression analysis for the objective functions minimize Total Project Delays, Total Weighted Project Delays, and Percent Increase In The Longest Critical Path, respectively.

TABLE 4. *Summary Statistics For Sample Independent Variables*

[60 multiproject test problems]

Variable	Mean (average)	Standard deviation	Variable	Mean (average)	Standard deviation	Variable	Mean (average)	Standard deviation
$\sum$DUR	1,395.43	621.23	PDENSITY–T	0.47	0.07	VA–CON	0.08	0.02
$\overline{\text{X}}$DUR	7.49	0.40	PCTFREESLK	0.09	0.02	MAXCON–TM	0.04	0.01
NSLACK	70.07	31.19	PDENSITY–F	0.82	0.05	MAXOFACT	11.07	8.23
PCTSLACK	0.38	0.03	$\overline{\text{X}}$%DEMAND	0.18	0.02	MAXUFACT	0.58	0.16
TOTSLACK–R	8.62	4.39	MINUTIL	0.11	0.06	MAXCON–ALL	0.24	0.13

Twelve independent variables given in section IV were omitted from the regression equations in Tables 6, 7, and 8 because of the multicolinearity which would have otherwise been present (several of the independent variables are partially derived from others). Variables eliminated because of zero-order correlations exceeding 0.90 with another variable left in the model include: NPROJ, NNODE, NARC, ΣDUR, T–DENSITY, ΣCPL, MAXCPL, ΣSLACK, PDENSITY–T, PDENSITY–F, MINCON–TM, and MAXCON–ALL.

In order to not bias the objective functions of minimizing Total Project Delays and Total Weighted Project Delays, a record was kept of the delays and weighted delays in the projects as of a specific

TABLE 5. *Factor Loadings Exceeding 0.70 In. Absolute Value—Varimax Rotation**

[60 multiproduct scheduling problems]

Variable	Factor loadings (Parentheses indicate negative loading)					
	F–1	F–2	F–3	F–4	F–5	F–6
NDUMMY	0.76					
$\bar{X}$CPL			(0.82)			
VA–CPL			(0.70)			
NSLACK	0.82					
$\bar{X}$SLACK					(0.89)	
TOTSLACK–R	0.76					
$\bar{X}$SLACK–R					(0.91)	
ΣFREESLK	0.70					
NFREESLK	0.78					
MINUTIL				0.85		
$\bar{X}$UTIL		-0.88				
MAXUTIL		0.87				
$\bar{X}$DMND	0.82					
MINCON	(0.70)					
$\bar{X}$CON	(0.89)					
VA–CON						(0.82)
$\bar{X}$CON–TM	(0.79)					
MAXCON–TM			0.72			
TOTOFACT				(0.83)		
MINOFACT		(0.83)				
MAXOFACT				(0.91)		
TOTUFACT		0.82				
MAXUFACT		0.92				
$\bar{X}$OVER			(0.84)			
MAXOVER			(0.92)			
$\bar{X}$UNDER		0.81				
MAXUNDER		0.91				
NCON–ALL	(0.70)					
$\bar{X}$CON–ALL	(0.94)					
VA–CON–ALL	(0.75)					

*The six factors computed retain 71 percent of the information accounted for by all independent variables.

time in each schedule span (usually about 75 percent of the length), and the objective functions were evaluated as of this time. This removed bias which would have otherwise been present as projects entered the completion phase and the simultaneous demands for resources declined.

Tables 6, 7, and 8 reveal some rather interesting relationships for the data examined. For example, specific independent variables remain in (stepdown regression) each regression equation with differing solution techniques (heuristics), indicating that a statistical relationship does exist between problem structures and the method employed to solve each problem. Note also that some of the project parameters—especially those assessing relationships between resource requirements and resource availabilities—are significant in a majority of the regression equations, indicating their importance in scheduling project activity for all solution procedures. Additionally, from a comparison of Tables 6, 7,

TABLE 6. *Multiple Regression*

[60 hypothetical multiproject

Project summary measures	Time and network based computed prior to critical path analysis					Time and network based computed subsequent to critical path analysis											Based on										
	NDUMMY	$\bar{X}$DUR	VA-DUR	$\bar{X}$DENSITY	COMPLEXITY	$\bar{X}$CPL	VA-CPL	NSLACK	PCTSLACK	$\bar{X}$SLACK	TOTSLACK-R	$\bar{X}$SLACK-R	ΣFREESLK	NFREESLK	PCTFREESLK	$\bar{X}$FREESLK	MIN%DEMAND	$\bar{X}$%DEMAND	MAX%DEMAND	MINUTIL	$\bar{X}$UTIL	MAXUTIL	XDMND	MINCON	XCON	MAXCON	VA-CON
Scheduling heuristics: LTF		+	+			−	+	+			−	+	+			−	+	+			+	-		−	+	+	−
GRD						−							+								+	−				+	
GRRD								+										+								+	−
SIO			−			−		+		−		+		−									+	−	+	+	
GRU	+	+	+			−	−	+			−	+	+			−	+	+	−		+			+	+	+	−
LFT		+	+		−	−				−		+	+			−	+	+	−			+			+	+	−
MJP						−	−	+							+		+						+		+		
RAN			+					+	−																	+	

*Each net regression coefficient significant at the 95 percent level.
(+) indicates positive net regression coefficient.
(−) indicates negative net regression coefficient.

TABLE 7. *Multiple Regression Results*

[60 hypothetical multiproject

Project summary measures	Time and network based computed prior to critical path analysis					Time and network based computed subsequent to critical path analysis											Based on												
	NDUMMY	$\bar{X}$DUR	VA-DUR	$\bar{X}$DENSITY	COMPLEXITY	$\bar{X}$CPL	VA-CPL	NSLACK	PCTSLACK	$\bar{X}$SLACK	TOTSLACK-R	$\bar{X}$SLACK-R	ΣFREESLK	NFREESLK	PCTFREESLK	$\bar{X}$FREESLK	MIN%DEMAND	$\bar{X}$%DEMAND	MAX%DEMAND	MINUTIL	$\bar{X}$UTIL	MAXUTIL	$\bar{X}$DMND	MINCON	$\bar{X}$CON	MAXCON	VA-CON	$\bar{X}$CON-TM	MAXCON-TM
Scheduling heuristics: LTF			+							−		+	+			−									+	+	−		
GRD		+	+			−	+				−	+	+			−	+	+	−	+	+	−		+	+	+	−	−	+
GRRD		+	+			−				−		+	+			−	+	+	−			+			+	+	−	−	+
SIO	−		+							−		+	+			−		+						+	+	+	−		
GRU			+			−	−				-	+	+	+		−	+	+	−		+				+	+	−		
LFT		+	+		−	−				−		+	+			−	+	+	−	+		+				+	−	−	+
MJP																	+					+							
RAN													+			−		+								+			

*Each net regression coefficient significant at the 95 percent level.
(+) indicates positive net regression coefficient.
(−) indicates negative net regression coefficient.

Results for Total Project Delays

scheduling problems*]

resource usage																		Summary statistics and regression results			
																		Summary statistics		Regression results	
$\bar{X}$CON-TM	MAXCON-TM	VA-CON-TM	MINCON-ALL	$\bar{X}$CON-ALL	VA-CON-ALL	TOTOFACT	MINOFACT	MAXOFACT	TOTUFACT	MINUFACT	MAXUFACT	$\bar{X}$OVER	MINOVER	MAXOVER	$\bar{X}$UNDER	MINUNDER	MAXUNDER	Average total project delays	Standard deviation	Standard error of estimate about regression	R^2
−	+		−	−	+		−					+		−				390.82	232.05	64.08	0.96
				−						+		+					+	359.60	211.81	67.64	0.91
									+			+						358.93	204.89	61.46	0.92
		−				+		−				+			+			326.73	200.52	66.77	0.92
−				−		+	−	−		−		+					+	356.30	213.22	45.16	0.97
−	+	−	−	−		+	−	−				+	−		+			372.95	224.28	55.76	0.96
	+	−				+		−	−		+			+	+			337.70	203.58	74.21	0.90
									+			+						373.60	222.53	76.32	0.89

for Total Weighted Project Delays

scheduling problems*]

resource usage																Summary statistics and regression results			
																Summary statistics		Regression results	
VA-CON-TM	MINCON-ALL	$\bar{X}$CON-ALL	VA-CON-ALL	TOTOFACT	MINOFACT	MAXOFACT	TOTUFACT	MINUFACT	MAXUFACT	$\bar{X}$OVER	MINOVER	MAXOVER	$\bar{X}$UNDER	MINUNDER	MAXUNDER	Average weighted delays	Standard deviation	Standard error of estimate about regression	R^2
					−		+			+						317,962	206,664	71,183	0.90
	−	−	+	+	−	−	+	+	−	+		−			+	287,495	183,392	36,005	0.98
	−	−		+	−	−		−		+						287,970	177,508	35,611	0.97
		−			−		+			+						271,129	181,779	62,720	0.91
		−		+	−	−	+	−		+	+				+	289,576	183,628	40,940	0.97
−	−	−		+	−	−	+			+	−				+	302,709	194,513	42,184	0.97
				+	−	−	+					+				275,561	178,562	81,843	0.82
							+			+						304,636	194,579	72,888	0.88

TABLE 8. *Multiple Regression Results for*

[60 hypothetical multiproject

Project summary measures	Time and network based computed prior to critical path analysis					Time and network based computed subsequent to critical path analysis											Based on												
Scheduling heuristics	NDUMMY	$\bar{X}$DUR	VA–DUR	$\bar{X}$DENSITY	COMPLEXITY	$\bar{X}$CPL	VA–CPL	NSLACK	PCTSLACK	$\bar{X}$SLACK	TOTSLACK–R	$\bar{X}$SLACK–R	ΣFREESLK	NFREESLK	PCTREESLK	$\bar{X}$FREESLK	MIN%DEMAND	$\bar{X}$%DEMAND	MAX%DEMAND	MINUTIL	$\bar{X}$UTIL	MAXUTIL	$\bar{X}$DMND	MINCON	$\bar{X}$CON	MAXCON	VA–CON	$\bar{X}$CON–TM	MAXCON–TM
LTF								+			−							−	+	+		+	+	−				−	
GRD										+		−					+	−			−	+			−				
GRRD						+	+														−	+	+						
SIO				+							−	+		+	−		+					+		−			−		
GRU		+	+														+		−			+					−	−	+
LFT										+	−							−	+			+		−				−	
MJP		+	+				+	−		−		+		+	−				−			+		−	+			−	
RAN						+	+				−	+	+			−	+			+		+			+		−	−	+

*Each net regression coefficient significant at the 95 percent level.
(+) indicates positive net regression coefficient.
(−) indicates negative net regression coefficient.

and 8, which of the parameters remain in solution is a function of the objective function being evaluated, indicating that the choice of a scheduling rule is also dependent upon the desired result of scheduling effort.

A comparison of the columns "Standard Deviation" and "Standard Error of Estimate About Regression" and an examination of the R^2 values provide some indication of the efficacy of our procedure. In general, no less than 82 percent of the variability in results obtained using any heuristic procedure can be explained by the regression models developed.

An experiment to further assess the efficacy of these regression models might consist of the following. For each of the objective functions evaluated, rank the heuristics on the lowest mean value attained, the next lowest mean value attained, etc. This gives an order in which to select a heuristic procedure for solving problems, not considering unique problem structures. Alternatively, predict (using the models developed) the heuristic rule which will likely produce the lowest value of the objective function, the next lowest value, etc. for *specific* problems. Then, one, two, three, . . . of the heuristic rules can be employed to schedule project activity, and the following question addressed: "What is the expected improvement (if any) in using the rules ranked by the regression equations on each problem as opposed to selecting the rules on the basis of lowest average results obtained?"

Figure 3 presents data in this regard. For example, if a heuristic is selected to minimize project delays on the basis of lowest mean value obtained in solving these multiproject scheduling problems (in this case, the SIO heuristic), the amount of total project delay will be 16 percent above what could be obtained by solving these same problems with all rules and then selecting the schedule with the minimum amount of total delay. If instead, each problem is solved using the heuristic *projected best*

Percent Increase in Longest Critical Path Length

scheduling problems*]

resource usage																Summary statistics and regression results			
																Summary statistics		Regression results	
VA-CON-TM	MINCON-ALL	$\overline{X}$CON-ALL	VA-CON-ALL	TOTOFACT	MINOFACT	MAXOFACT	TOTUFACT	MINUFACT	MAXUFACT	$\overline{X}$OVER	MINOVER	MAXOVER	$\overline{X}$UNDER	MINUNDER	MAXUNDER	Average percent increase in longest critical path length	Standard deviation	Standard error of estimate about regression	R^2
		+			+		−	−		−			+			197.23	75.46	8.27	0.99
		+	−			−				−						210.88	79.30	13.84	0.98
					+			−		−						207.29	78.73	14.79	0.97
												+			−	211.04	82.84	16.41	0.96
																212.69	81.84	16.99	0.96
		+		−	+							−	+			196.45	74.93	8.27	0.99
									−		+				+	223.55	88.70	15.74	0.98
	−							−				−				209.68	78.70	13.72	0.98

on increase in total project delays, then the solution on the average will be only 8 percent above what could be obtained by using *all* heuristic procedures. Using only *one* heuristic sequencing rule to solve a given problem (which might usually be the case), use of the regression equations results in an *expected* improvement of 8 percent in this objective function value (16–8 percent). Naturally, as the number of solution techniques examined increases, the relative advantage of using the regression models developed decreases.

The use of the regression equations for minimizing Percent Increase in Longest Critical Path Duration generally does not produce a substantial improvement over using the rules based on lowest average value obtained, as indicated in Figure 3. The appropriateness of this particular objective function is certainly open to question, however, as discussed earlier. And results in minimizing Total Weighted Project Delays using the regression models are encouraging, although the differences in results obtained are not as great as in the minimization of Total Project Delays.

In order to further test the approach reported, relevant project scheduling literature was reviewed to gather project data from other sources for analysis. The most extensive set of data obtained consists of 83 *single* project scheduling problems generated by Davis [2] for use in testing his bounded enumeration algorithm. Davis used resource and network parameters nearly identical to those used by Johnson [9] is his investigation of the single constrained-resource, project scheduling problem. Johnson based his network generation routines on experience gained from two sources. He examined more than 50 CPM networks submitted for processing at one computing center. He also examined networks that have appeared elsewhere in various case studies. Briefly, each project consists of either 22 or 27 activities (nodes), and each activity can require fixed amounts of three scarce resources. Pertinent statistics for

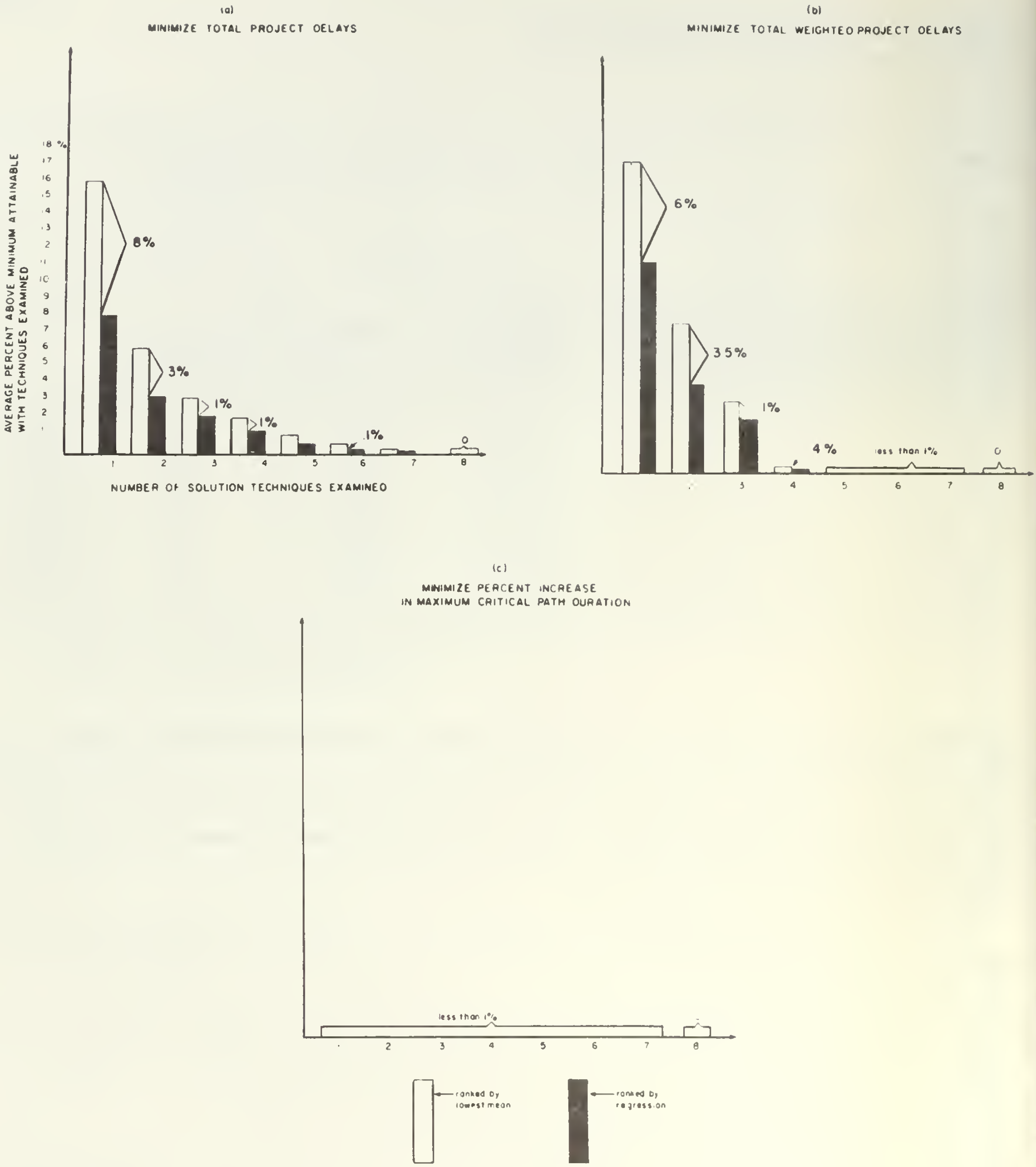

FIGURE 3. Percent increase in minimum heuristic results obtained—multiproject data.

many of the independent variables given in section IV are presented in Table 9 for the 83 problems. A factor analysis performed to combine the independent variables into factors is given in Table 10. Because single project data are examined, only the objective of minimizing the percent increase in critical path duration is assessed. As noted by Davis [4], this criteria also minimizes the delay in the completion of the single project.

Table 9. *Summary Statistics For Sample Independent Variables*

[83 single project test problems]

Variable	Mean (average)	Standard deviation	Variable	Mean (average)	Standard deviation	Variable	Mean (average)	Standard deviation
ΣDUR	73.48	12.51	PDENSITY-T	0.52	0.10	VA-CON	0.001	0.002
$\bar{X}$DUR	2.99	0.38	PCTFREESLK	0.27	0.09	MAXCON-TM	0.04	0.009
NSLACK	14.85	2.44	PDENSITY-F	0.74	0.06	MAXOFACT	0.38	0.16
PCTSLACK	0.60	0.07	$\bar{X}$%DEMAND	0.90	0.004	MAXUFACT	0.34	0.11
TOTSLACK-R	2.57	0.93	MINUTIL	0.78	0.13	MAXCON-ALL	0.32	0.07

Table 10. *Factor Loadings Exceeding 0.70 In Absolute Value—Varimax Rotation**

[83 single project scheduling problems]

Variable	Factor loadings (Parentheses indicate negative loading)					
	F-1	F-2	F-3	F-4	F-5	F-6
NARC				0.92		
ΣDUR					(0.75)	
VA-DUR					(0.70)	
COMPLEXITY				0.96		
ΣCPL					(0.91)	
NSLACK			(0.70)			
PCTSLACK			(0.85)			
TOTSLACK-R			(0.71)			
ΣFREESLACK	0.70					
PCTFREESLK				0.93		
$\bar{X}$%DEMAND		0.86				
MINUTIL	(0.91)					
MINCON	(0.86)					
VA-CON						(0.94)
MINCON-TM	(0.75)					
MAXCON-TM	(0.70)					
VA-CON-TM						(0.90)
MINOFACT	0.75					
MAXOFACT	0.81					
TOTUFACT	(0.86)					
MAXUFACT	(0.79)					
MAXOVER					(0.70)	
MINUNDER	(0.82)					
MAXUNDER	(0.74)					
MAXCON-ALL						(0.70)

*The six factors computed retain 83 percent of the information accounted for by all independent variables.

TABLE 11. *Multiple Regression Results for Percent Increase in Critical Path Duration*

[83 hypothetical single project scheduling problems*]

		Time and network based computed prior to critical path analysis				Time and network based computed subsequent to critical path analysis							Based on resource usage																		Summary statistics and regression results			
																															Summary statistics		Regression results	
	Project summary measures	ΣDUR	VA-DUR	$\bar{X}$DENSITY	COMPLEXITY	$\bar{X}$CPL	NSLACK	PCTSLACK	TOTSLACK-R	ΣFREESLK	PCTFREESLK	PDENSITY-F	$\bar{X}$%DEMAND	MINUTIL	MINCON	VA-CON	MINCON-TM	MAXCON-TM	VA-CON-TM	VA-CON-ALL	MINOFACT	MAXOFACT	TOTUFACT	MINUFACT	MAXUFACT	MINOVER	MAXOVER	MINUNDER	MAXUNDER	Average percent increase in critical path duration	Standard deviation	Standard error of estimate about regression	R^2
Scheduling heuristics	LTF	+				−			−			+		+		−	−		+	+		+	−		+					27.23	23.93	6.64	0.93
	GRD												−	+				+		+	+									36.31	26.88	10.46	0.86
	RSM											+	−	+		−		+	+	+		+		+						28.51	22.78	6.48	0.93
	SIO	+			−		−	+						+		−		+				+					−			38.31	21.77	9.48	0.83
	GRU	+							−					+								+	−		+					35.78	21.78	7.30	0.90
	LFT						+	−	−					+		−		+		+		+	−		+					28.37	21.93	7.32	0.90
	MJP												−	+				+		+		+			−	+	−			39.20	23.39	8.39	0.88
	RAN	+	−			−	−	+					−	+			−	+				+								34.05	24.27	8.56	0.89

*Each net regression coefficient significant at the 95 percent level.

(+) indicates positive net regression coefficient

(−) indicates negative net regression coefficient

The regression results obtained using this single project data are given in Table 11. Several of the independent variables found in the previous regression results do not appear in this table because of the higher number of zero-order correlations exceeding 0.90. Two independent variables, PDENSITY–F and MINCON–TM, not present in the previous regression equations, remained in for this data.

Despite the great dissimilarity of data types examined (single vs. multiple projects; three vs. 13 resource categories; etc.), similar tendencies can be noted in the results reported in Table 8 and Table 11. For example, there is a tendency for the heuristics LTF and LFT to perform the best on the criteria examined, and for the heuristics MJP and SIO to perform relatively worse than a rule which resolves resource conflicts at random in both sets of data. Many of the independent variables which remained in the regression equations for the previous results are also present for the Davis data even though several variables were eliminated from the candidate variables list because of problems with multicolinearity which would have otherwise developed. Additionally, parameters based on resource utilization are significant in every regression equation for both sets of data: the minimum resource utilization parameter is significant in the Davis problems; the maximum resource utilization parameter is significant for the multiproject data.

Even the above discrepancy in the MINUTIL vs. MAXUTIL parameter being significant in the regression equations for these two important classes of problems can be resolved. The difference in values of minimum and maximum resource utilization for the Davis single project data (averaged across all problems) is $0.90-0.78=0.12$; for the multiproject scheduling data it is $1.67-0.11=1.56$, a much larger difference. Moreover, the parameter MAXUTIL is highly correlated with MAXCON–TM in the Davis data ($\rho=0.96$), so that the MAXUTIL parameter was not permitted to enter as an independent variable. The MAXCON–TM parameter is significant at the 0.95 level in six of the eight equations reported.

The R^2 values for each regression equation for this second group of data range from a low of 0.83 to a high of 0.93, indicating again that a statistical relationship does exist between the parameters included and the objective of scheduling effort.

VI. AN IMPROVED PROGRAM FOR PROJECT SCHEDULING

Given the improvements reported with the use of prediction equations described, advantage can be gained by constructing a program incorporating *all* heuristic procedures. Then, heuristic rule(s) can be selected from those available based upon unique problem characteristics for solving individual problems. Figure 4 is a simplified flow chart of such a program. Figure 3 provides some indication of the advantages of such an approach over a similar program which incorporates only one solution technique and in which no attempt is made to determine beforehand the solution procedure to use.

VII. SUMMARY AND CONCLUSIONS

Scheduling problems often exhibit unique characteristics which make them more amenable to solution with a particular procedure. The advantages of analyzing problem structure and *then* choosing a technique for solving it can be significant. For the data examined herein, they are as indicated in Figure 3.

For certain other problems, the choice of a scheduling technique (rule) is relatively unimportant—nearly all will produce schedules with the same or nearly the same objective function value. These

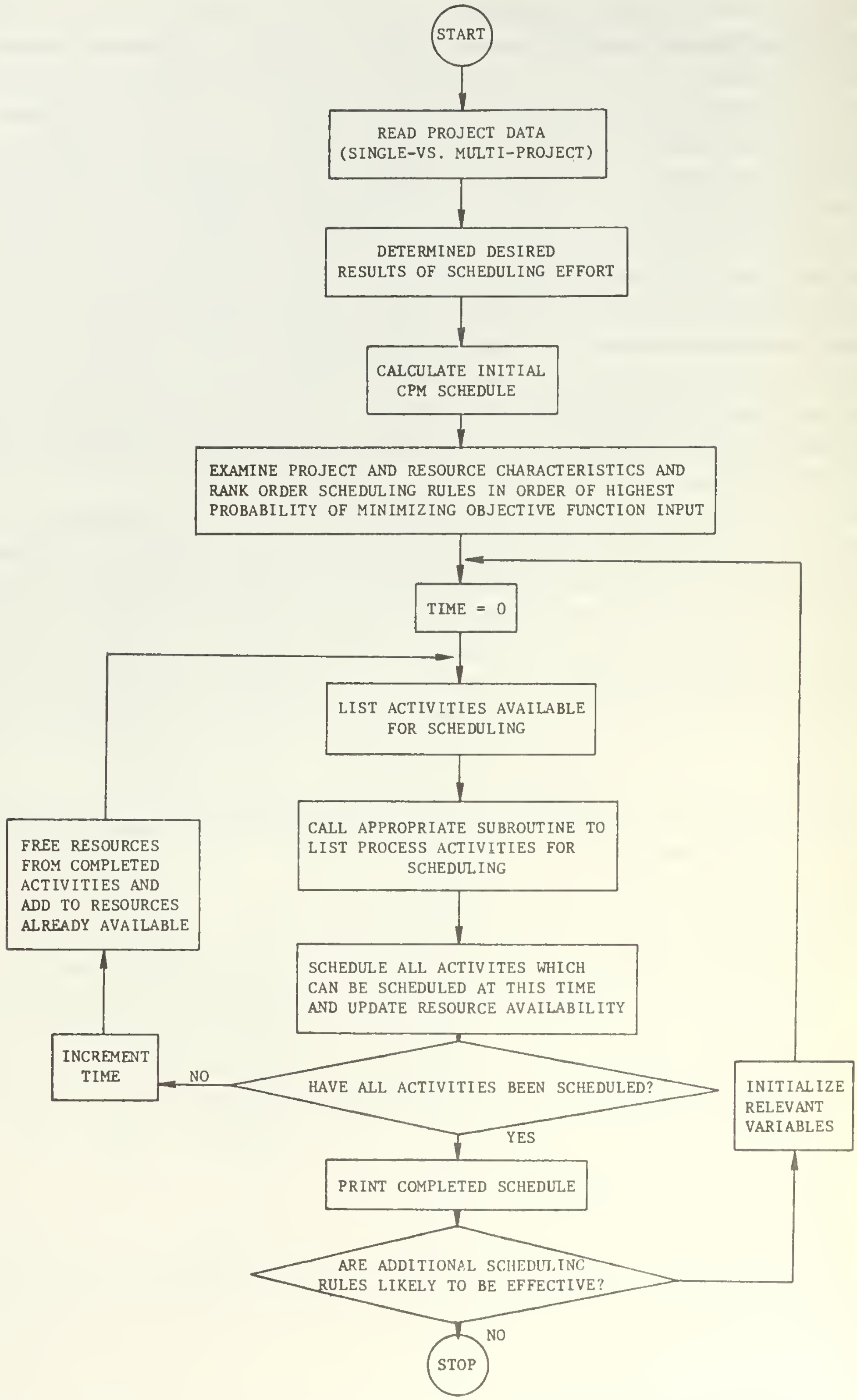

FIGURE 4. Flow chart of scheduling program which exploits problem structure.

problems generally consist of projects in which the initial obstruction as determined by an all early start schedule is low. Several indices of "resource constrainedness" were also described which can contribute to the masking of heuristic performance.

Subsequent to the computation of the regression models reported, attempts were made using step-wise regression and multiple correlation to determine characteristics of problems which would tend to mask the effectiveness of the heuristic procedures employed. In general, it was found that greater differences in heuristic performance are likely to occur for those projects or those project sets in which: 1) initial obstruction is high; 2) projects possess relatively large amounts of and occurrences of free float; and 3) certain resource constrainedness measures are low. Naturally, for example, as the amount of obstruction in a problem decreases, few resolutions of sequencing conflicts have to be made, and the results obtained with different methods are similar.

The following guidelines are useful for selecting a procedure to use in scheduling multiproject activity to minimize total project delays: The Shortest Imminent Operation heuristic generally performs the best, particularly when the variance in activity duration is small (but is not zero!). Further, projects possessing a small proportion of activities of rather long duration are not effectively scheduled with this rule. This is because, of course, the larger duration activities have a tendency to be postponed too long, resulting in greater total project delays. (These results agree with those reported elsewhere for the job shop.) Working on as many jobs (activities) at a time as is possible can be effective in reducing total project delays when the SIO rule has been discarded because of high variability in activity duration, although this rule can be extremely dysfunctional when trying to minimize other measures of heuristic performance. And in general, the Least Total Float and the Late Finish Time heuristics are comparatively effective on those problems in which the variability in activity duration and the average free slack per activity is high. We would expect these latter two sequencing rules to exhibit similar performance, since for each activity they differ from one another only by activity duration.

The results reported also demonstrate the necessity of determining beforehand the desired results of scheduling effort. In minimizing project makespan, for example, a combination of the Least Total Float or the Late Finish Time rules resulted in 44 out of 60 minimum makespan schedules for the multiproject data, and in 65 out of 83 minimum makespan schedules for the single project data. There was additionally little discernible difference in results reported using these two procedures to minimize project makespan for multiproject activity. By contrast, the Shortest Imminent Operation heuristic produced only 18 out of 60 schedules resulting in minimum total project delays, and the variation in results obtained with different rules was much more significant in a majority of the problems examined. Hence, the determination of the most likely rule to schedule project activity effectively is not only a function of problem characteristics, but is also a function of the desired results of scheduling effort. For minimizing project makespan, one is fairly "safe" in using the Least Total Float or the Late Finish Time rule; for other criteria of scheduling, regression analysis becomes an effective tool in selecting the appropriate sequencing discipline to employ.

Perhaps an even greater use to be gained from the results reported in this paper is in the area of training project schedulers and program managers. Using an interactive system, for example, program managers can examine project and resource characteristics and through projections of their own and the actual results of scheduling, learn which sequencing rules are likely to be effective under various problem characteristics. This man/machine interaction should also enhance our ability to learn more of the role between problem structure and heuristic performance in scheduling multiproject activity.

ACKNOWLEDGMENT

The author would like to acknowledge the computer programming assistance provided by Mr. Mike Neary in the preparation of this paper. Mr. Neary also provided many worthwhile ideas and comments on earlier drafts of the paper. Funds for completing the project were provided for by an NSF Institutional Grant to The Pennsylvania State University. Additionally, several smaller grants were received from the Center for Research, College of Business Administration, at Penn State.

REFERENCES

[1] Brand, J. D., W. L. Meyer, and L. R. Shaffer, "The Resource Scheduling Problem in Construction," Civil Engineering Studies, Report No. 5, Urbana, Illinois: Department of Civil Engineering, University of Illinois (1964).

[2] Davis, E. W., "An Exact Algorithm for the Multiple Constrained-Resource Project Scheduling Problem," unpublished Ph.D. Thesis, Yale University (1969).

[3] Davis, E. W., "Project Scheduling Under Resource Constraints—Historical Review and Categorization of Procedures," AIIE Transactions, *5*, 297–313 (1973).

[4] Davis, E. W., "Project Summary Measures and Constrained Resource Scheduling," Working Paper HBS 73–34, Graduate School of Business Administration, Harvard University (1973).

[5] Davis, E. W. and G. E. Heidorn, "An Algorithm for Optimal Project Scheduling Under Multiple Resource Constraints," Management Science, *17*, B–803–816 (1971).

[6] Davis, E. W. and J. H. Patterson, "A Comparison of Heuristic and Optimum Solutions In Resource-Constrained, Project Scheduling," Working Paper HBS 73–15, Graduate School of Business Administration, Harvard University, Boston, Mass. (1973). (Also to appear in Management Science)

[7] Davies, Mary E., "New Criterion For Resource Scheduling," Transportation Engineering Journal, *99*, No. TE4, 741–755 (1973).

[8] Fendley, L. G., "Toward the Development of a Complete Multiproject Scheduling System," Journal of Industrial Engineering, *19*, 505–515 (1968).

[9] Johnson, Thomas J. R., "An Algorithm for the Resource-Constrained, Project Scheduling Problem," Unpublished Ph.D. Thesis, Massachusetts Institute of Technology (1967).

[10] Mize, J. H., "A Heuristic Scheduling Model for Multi-Project Organizations," Unpublished Ph.D. Thesis, Purdue University (1964).

[11] Pascoe, T. L., "An Experimental Comparison of Heuristic Methods for Allocating Resources," Unpublished Ph.D. Thesis, Cambridge University (1965).

[12] Patterson, James H., "Alternate Methods of Project Scheduling With Limited Resources," Nav. Res. Log. Quart., *20*, 767–784 (1973).

[13] Patterson, James H. and Walter D. Huber, "A Horizon-Varying, Zero-One Approach to Project Scheduling," Management Science, *20*, 990–998 (1974).

[14] Patterson, James H., "An Empirical Investigation of Alternate Methods for Scheduling Multiple Projects With Limited Resources," *Proceedings, Third International Congress on Project Planning by Network Techniques*, Stockholm, Sweden, 294–309 (May, 1972).

[15] Wiest, J. D., "Some Properties of Schedules for Large Projects with Limited Resources," Operations Research, *12*, 395–418 (1964).

[16] Wiest, J. D., "A Heuristic Model for Scheduling Large Projects with Limited Resources," Management Science, *13*, 359–377 (1967).

[17] Wiest, J. D., "The Scheduling of Large Projects with Limited Resources," Unpublished Ph.D. Thesis, Carnegie Institute of Technology (1963).

SCHEDULING TO MINIMIZE THE WEIGHTED SUM OF COMPLETION TIMES WITH SECONDARY CRITERIA

Richard N. Burns
Australian National University
Canberra, Australia

ABSTRACT

A result of Smith previously published in this journal [3], on the use of secondary criteria in scheduling problems, is shown to be incorrect and a counter example is presented.

Heck and Roberts [2] suggested that their paper would be extended in the same way Smith's algorithm was. A new algorithm is given that converges to a local optimum for both problems.

1. BACKGROUND

The problem of minimizing the sum of weighted completion times subject to a secondary constraint has long been considered to be similar to minimizing the sum of completion times subject to a constraint. Smith [3] 1956 presented an algorithm for the n-job, 1-machine case where he minimized the weighted completion times subject to the constraint that all jobs were completed by their due dates. Recently Heck and Roberts [2] presented an algorithm for minimizing the sum of completion times subject to not increasing the maximum tardiness calculated by the due date sequence. They claimed that their result could be extended to the problem of weighted completion times in a manner similar to Smith's algorithm. Such is not the case since Smith's algorithm did not find the optimum sequence.

Using the notation of Conway, Maxwell and Miller [1], let job i have process time p_i, due date d_i, weight a_i and completion time c_i. The problem under consideration is to find a sequence to

$$\text{minimize} \sum_{i=1}^{n} a_i c_i \tag{1}$$

$$\text{subject to} \quad \sum_{j=1}^{i} p_j - d_i \leq T \qquad \text{for } i = 1, 2, \ldots, n \tag{2}$$

For $T=0$ in (2) the problem is the one considered by Smith [3] and for $T=$ maximum tardiness the problem is the one of Heck and Roberts [2].

2. A COUNTER EXAMPLE TO SMITH'S ALGORITHM

The method given by Smith to solve (1) subject to (2) with $T=0$ is as follows: "If all jobs can be

completed by their due dates, an order which minimizes the weighted sum may be obtained. This order has its last job one with the largest value of p_i/a_i from those with due date as large as the total processing time of all jobs."

Consider the following example:

job	1	2	3
p_i	4	3	2
d_i	8	9	10
a_i	1	4	3
p_i/a_i	4	3/4	2/3

Using Smith's algorithm, job 2 would be placed last, since job 2 and 3 both have due dates greater than or equal to the total process time and $3/4 > 2/3$. Considering the remaining two jobs, again both job 1 and 3 may be last and $4 > 2/3$. Therefore the optimum sequence is 3, 1, 2 with $\sum_{i=1}^{3} a_i c_i = 48$. However, the sequence 2, 1, 3 also has all jobs completed by their due date and $\sum_{i=1}^{3} a_i c_i = 46$. Hence Smith's algorithm does not give the correct sequence.

When considering job k to be placed last it is necessary to check that the resulting sequence satisfies (2). Hence a more precise statement of Smith's algorithm would be as follows:

ALGORITHM 1: (a) Order the jobs in the order of increasing due dates. It is assumed that all jobs are on time by this schedule.

(b) Place job k last where

$$\frac{p_k}{a_k} \geqslant \frac{p_i}{a_i} \quad \text{for all } i \text{ such that the resulting sequence will satisfy (2).}$$

(c) Reduce n by 1 and return to (b) until all jobs have been sequenced by this method.

3. LOCAL OPTIMUM

Two sequences S and S' are said to be adjacent if one can be formed from the other by a single interchange of two jobs. A sequence is feasible if and only if it satisfies (2).

DEFINITION: A sequence is a local optimum for (1) subject to (2) if $\sum_{i=1}^{n} a_i c_i$ is less than or equal to the sum of weighted completion times of all feasible adjacent sequences.

In order to check for a local optimum a method of comparing two adjacent sequences will now be developed.

Let S be a sequence with the jobs numbered $1, 2, \ldots, n$ and let S' be an adjacent sequence having jobs k and L interchanged. Without loss of generality we can assume that job L appears after job k in the sequence S.

For $i \leq k-1$ and for $i \geq L$,

$$a_i c_i = a_i' c_i'.$$

For $i = k, k+1, \ldots, L-1$,

$$c_i' = c_i + (p_L - p_k).$$

Therefore

$$\sum_{i=1}^{n} a_i c_i - \sum_{i=1}^{n} a_i' c_i' = a_k c_k + a_L c_L - a_k c_L - a_L c_k + (p_k - p_L) \sum_{i=k+1}^{L} a_i.$$

The following lemma follows directly from the definition of a local optimum and equation (3).

LEMMA 1: A sequence S is a local optimum if and only if for all feasible sequences differing from S by having jobs L and k interchanged with $k < L$ the expression

$$a_L c_L + a_k c_k - a_k c_L - a_L c_k + (p_k - p_L) \sum_{i=L+1}^{k} a_i \leq 0. \tag{4}$$

For the special case where all the weights are one then equation (3) becomes

$$\sum_{i=1}^{n} c_i - \sum_{i=1}^{n} c_i = (p_k - p_L)(L-k). \tag{5}$$

Note that for the unweighted case the difference between two adjacent sequences is only a function of the two terms being interchanged and their distance apart. Lemma 1 would have an equivalent form for the unweighted case where equation (5) would be substituted for equation (4). The proof that Algorithm 1 converges to a local optimum (for the unweighted case) is straight forward and in fact was done correctly by Smith as he essentially developed equation (5).

For the weighted case the difference between adjacent sequences is a function of the jobs between the two interchanged jobs. In the papers by both Smith and Heck and Roberts a very special case of equation (3) was used. When $L = k+1$ equation (3) becomes the following.

$$\begin{aligned}\sum_{i=1}^{n} a_i c_i - \sum_{i=1}^{n} a_i' c_i' &= a_k c_k + a_{k+1} c_{k+1} - a_k c_{k+1} - a_{k+1} c_k + (p_k - p_{k+1}) a_{k+1} \\ &= a_k \sum_{i=1}^{k-1} p_i + a_k p_k + a_{k+1} \sum_{i=1}^{k-1} p_i + a_{k+1} p_k + a_{k+1} p_{k+1} \\ &\quad - a_k \sum_{i=1}^{k-1} p_i - a_k p_k - a_{k+1} \sum_{i=1}^{k-1} p_i - a_{k+1} p_k + a_{k+1} p_k - a_{k+1} p_{k+1}\end{aligned}$$

$$\sum_{i=1}^{n} a_i c_i - \sum_{i=1}^{n} a_i c_i = a_{k+1} p_k - a_k p_{k+1}. \tag{6}$$

Equation (6) is the test used in step (b) of Algorithm 1.

LEMMA 2: Algorithm 1 (Smith) converges to a constrained stationary sequence S having the property that $\sum_{i=1}^{n} a_i c_i$ for S is less than or equal to the sum of weighted completion times for the subset of feasible adjacent sequences formed by interchanging only pairs of jobs that are adjacent in S.

PROOF: Let the jobs in S be numbered in ascending order, $1, 2, \ldots, n$. Assume the Lemma is false. Then there exists S' a feasible sequence adjacent to S with

$$\sum_{i=1}^{n} a_i c_i - \sum_{i=1}^{n} a_i' c_i' > 0,$$

and S and S' differ only in the k and $k+1$ st position. By (6)

$$\frac{p_k}{a_k} > \frac{p_{k+1}}{a_{k+1}} \cdot$$

When Algorithm 1 was choosing the job for the $k+1$ position job k was rejected and later placed in position k. Since S' is also feasible, job k must have been eligible to go in the $k+1$ st position. By step (b) of Algorithm 1

$$\frac{p_{k+1}}{a_{k+1}} \geqslant \frac{p_k}{a_k}$$

which contradicts the result of the assumption.

The counter example presented shows that considering only interchanges between jobs that are adjacent in a sequence is not a sufficient condition for a sequence to satisfy Lemma 1. The following algorithm uses equation (2) to generate a local optimum.

ALGORITHM 2: 1. Schedule the jobs in increasing order of due date. (It is assumed that all jobs are completed by their due date. If such is not the case a solution to (1) subject to (2) with T equal to the maximum tardiness is straight forward.)

2. For $k=n-1, n-2, \ldots, 1$, find the first value of k satisfying the following three conditions.

$$T \geqslant \sum_{i=1}^{n} p_i - d_k \tag{7}$$

$$\text{If } p_n > p_k, \quad \text{then } c_j + p_n - p_k - d_j \leqslant T \quad \text{for } j=k+1, k+2, \ldots, n-1 \tag{8}$$

$$a_k c_k + a_n c_n - a_n c_k - a_k c_n + (p_k - p_n) \sum_{i=k+1}^{n} a_1 > 0.$$

If such a k is found, interchange job k and n. Reset n to its initial value if necessary and return to the beginning of step 2. If no k is found satisfying (7), (8) and (9) reduce n by 1. For $n \geqslant 2$ return to the beginning of step 2. For $n=1$ the optimum sequence is the current one.

THEOREM 1: Algorithm 2 converges to a local optimum.

PROOF: Every time step two is successful the sum of the weighted completion times is reduced. Since this can only happen a finite number of times the algorithm terminates. Conditions (7) and (8) determine which term in the sequence can be interchanged with job n and equation (9) is equivalent to having the right hand side of equation (3) greater than zero. Step 2, will only terminate after n has passed (7), (8) and (9) for all values of n from $1, 2, \ldots, n$. Hence by Lemma 1 the stationary point found by Algorithm 2 is a local optimum.

4. CONCLUSION

Algorithm 2 on the counter example would first find the stationary point that Smith's method stopped at, and then moves directly to the global optimum. It is not always the case that the local optimum found by Algorithm 2 is the global optimum.

Consider the following example:

job	1	2	3
p_i	4	3	6
d_i	8	13	14
a_i	1	4	9

Algorithm 2 would generate a local optimum sequence 1, 3, 2 with $\sum_{i=1}^{3} a_i c_i = 146$. The global optimum is 2, 1, 3 with $\sum_{i=1}^{3} a_i c_i = 136$. At present there is no simple way of finding a global optimum for permutation sequences.

REFERENCES

[1] Conway, R. W., W. L. Maxwell, and L. W. Miller, *Theory of Scheduling* (Addison-Wesley, Reading, Mass., 1967).

[2] Heck, H. and S. Roberts, "A Note on the Extensions of a Result on Scheduling with Secondary Criteria," Nav. Res. Log. Quart., *19*, 403–405 (1972).

[3] Smith, W. E., "Various Optimizers For Single-Stage Production," Nav. Res. Log. Quart., *3*, 59–66 (1956).

A CLASS OF DISTRIBUTIONS GENERATED FROM DISTRIBUTIONS OF EXPONENTIAL TYPE*

Regina C. Elandt-Johnson

Department of Biostatistics
University of North Carolina
Chapel Hill, North Carolina

ABSTRACT

A class of exponential type distributions with special exponential parameters is defined. It is assumed that the exponential parameters vary according to some (known) probability law. It has been shown in this paper that the compound distribution can be easily represented in form involving moment generating function of the mixing distribution. The results obtained in this paper provide an efficient and simple method of obtaining compound failure time distribution with known mixing distributions (uniform, exponential, gamma).

0. INTRODUCTION

One method (among others) of constructing a new distribution is to use the known parametric form of a distribution and allow one (or some) of the parameters to vary according to a specified probability law. The new distribution is called a *compound* distribution.

The theory of compound distributions is well known and frequently used in various scientific disciplines. In particular, it seems that this theory has useful applications in (industrial) reliability and (medical) survivorship analysis. The probability of surviving up to a certain time might be a function not only of the age of an individual, but also of other factors (parameters), $\bar{z}=\boldsymbol{\zeta}(\zeta_1, \zeta_2, \ldots, \zeta_k)$, which may not, in fact, be constants but vary from individual to individual and with time as well. For example, in survivorship analysis explanatory variables such as age, blood pressure, blood count, lipid levels, etc., etc. are sometimes observed. When they are not observed, their effects are nonetheless still present; the ζ's may be regarded as representing these effects. In the population, we then have a mixture (or compound) of distributions with ζ's as mixing variables.

It seems to be useful to have a simple and rapid method of obtaining a compound distribution—and this is the purpose of this article.

(a) It is shown in this paper that for a certain class of parental distributions, the compound distribution can be obtained from the moment generating function of the mixing distribution. This is another application of the mgf, besides that of obtaining the moments.

(b) Of some interest is the situation where the mgf is of exponential form. In this case the compound distribution can be factorized; the components due to parental and mixing distributions can be predicted.

The paper is not concerned with the use of the compound distributions in statistical inference.

*This work was supported by U.S. National Heart and Lung Institute contract NIH–NHLI–71–2243 from the National Institutes of Health.

The examples given in section 4 are mostly not new and are only used as illustrations.

To make the paper selfcontained, a few definitions will be introduced in the next section.

1. DISTRIBUTIONS OF EXPONENTIAL TYPE

For the purpose of this paper we find it convenient to define a special class of exponential type distributions in the following way.

DEFINITION 1: The distribution of a random variable (rv) X is of simple exponential type if its cumulative distribution function (cdf), $F_X(x)$, is of the form

$$F_X(x) = F_X(x;\eta;\boldsymbol{\alpha}) = \begin{cases} 0 & \text{for } x \leq x_0 \\ 1-\exp\left[-\eta u(x;\boldsymbol{\alpha})\right] & \text{for } x > x_0, \end{cases} \tag{1.1}$$

where $\eta(>0)$ is a "special" parameter, which we will call the exponential parameter, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_r)$ are "ordinary" parameters, and $u(x;\boldsymbol{\alpha})$ is an increasing function of x with

$$\begin{aligned} u(x;\boldsymbol{\alpha}) &\to 0 \qquad \text{as } x \to x_0, \\ u(x;\boldsymbol{\alpha}) &\to \infty \qquad \text{as } x \to \infty. \end{aligned} \tag{1.2}$$

We extend the definition to the multiparameter case.

DEFINITION 2: The distribution of a random variable X if of k-parameter exponential type if its cumulative distribution function, $F_X(x)$, is of the form

$$F_X(x) = F_X(x;\boldsymbol{\eta};\boldsymbol{\alpha}) = \begin{cases} 0 & \text{for } x \leq x_0 \\ 1-\exp\left[-\sum_{i=1}^{k}\eta_i u_i(x;\boldsymbol{\alpha}_i)\right] & \text{for } x > x_0 \end{cases} \tag{1.3}$$

and $\eta_i > 0$, $i = 1, 2, \ldots, k$. Here $\boldsymbol{\eta} = (\eta_1, \eta_2, \ldots, \eta_k)$ are exponential parameters and $u_i(x;\boldsymbol{\alpha}_i)$ are increasing functions of x such that

$$\begin{aligned} u_i(x;\boldsymbol{\alpha}_i) &\to 0 \qquad \text{as } x \to x_0, \\ u_i(x;\boldsymbol{\alpha}_i) &\to \infty \qquad \text{as } x \to \infty \end{aligned} \tag{1.4}$$

for $i = 1, 2, \ldots, k$.

Most lifetime distributions used in biology and reliability belong to the exponential type as defined in this article. For example, the Weibull distribution with cdf

$$F_X(x) = 1-\exp\left[-\eta(x-\xi)^\alpha\right], \qquad x > \xi, \qquad \eta > 0, \qquad \alpha > 0$$

is of exponential type with $u(x;\xi,\alpha) = (x-\xi)^\alpha$.

In this article we will discuss compound distributions of X when the parental distribution is of exponential type and the exponential parameters, $\eta_1, \ldots, \eta_k$ vary.

2. COMPOUND DISTRIBUTIONS OF X WHEN η IS THE MIXING PARAMETER

Let $F_X(x)$ be defined as in (1.1). The function

$$Pr\{X > x\} = 1 - F_X(x) = \bar{F}_X(x) \tag{2.1}$$

is called a *survival function.* (In reliability theory it is also called *reliability function*).

Consider η as a rv, with cdf

$$Pr\{\eta \leq y\} = P_\eta(y; \gamma) = P_\eta(y), \tag{2.2}$$

satisfying the conditions

$$\left.\begin{array}{rl} P_\eta(y) \to 0 & \text{as } y \to y_0 \\ 0 < P_\eta(y) < 1 & y_0 < y < \infty \\ P_\eta(y) \to 1 & \text{as } y \to \infty \end{array}\right\}. \tag{2.3}$$

(Here $\gamma = (\gamma_1, \ldots, \gamma_m)$ are parameters of the distribution (2.2)).

We will call $F_X(x)$ the *parental distribution*, and $P_\eta(y)$ the *mixing (or compounding) distribution.*

We assume that the moment generating function (mgf), $M_\eta(s)$, of the mixing distribution (2.2) exists. Since $\eta > 0$, all moments about zero of the distribution $P_\eta(y)$ are non-negative, and $M_\eta(s)$ is a nondecreasing function of s.

Let $G_X(x; \gamma, \alpha) = G_X(x)$ denote the cdf of the compound distribution, $F_X(x; \eta, \alpha) \underset{\eta}{\wedge} P_\eta(y; \gamma)$, and let $\bar{G}_X = 1 - G_X(x)$ be its survival function.

THEOREM 1: If the parental distribution of the rv X is of simple exponential type as defined in (1.1) and the exponential parameter η has a distribution (2.2) with mgf. $M_\eta(s)$, then the compound distribution of X, $G_X(x)$, is

$$G_X(x) = G_X(x; \gamma, \alpha) = \begin{cases} 0 & \text{for } x \leq x_0 \\ 1 - M[-u(x; \alpha)] & \text{for } x > x_0. \end{cases} \tag{2.4}$$

Proof is immediate. The survival function is

$$\bar{G}_X(x) = \int_{y_0}^{\infty} \exp\,[-u(x; \alpha)y]\,dP_\eta(y). \tag{2.5}$$

Put

$$s = -u(x; \alpha). \tag{2.6}$$

Thus the integral in (2.5) becomes

$$\bar{G}_X(x) = \int_{y_0}^{\infty} e^{sy}\,dP_\eta(y) = M_\eta(s) = M_\eta[-u(x; \alpha)], \tag{2.7}$$

provided that $M_\eta[-u(x; \alpha)]$ exists for $x > x_0$.

We find interesting to consider mixing distributions of η for which mgf has the form

$$M_\eta(s) = A(s, \boldsymbol{\delta})e^{\gamma s}, \tag{2.8}$$

where $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_m)$ are parameters. We now denote the parental distribution (1.1) by $F_{X;\eta}(x)$ to emphasize that it is a conditional distribution of X given η, and η can be considered as a random variable.

Corollary 1: If the parental distribution $F_{X;\eta}(x)$ is given by (1.1) and the mixing distribution $P_\eta(y; \gamma, \boldsymbol{\delta})$ has the mgf of the form (2.8), then the compound cdf of X, $G_X(x)$, is of the form

$$G_X(x) = 1 - A[-u(x; \boldsymbol{\alpha}); \boldsymbol{\delta}]\,[1 - F_{X;\gamma}(x)]. \tag{2.9}$$

Notice that $F_{X;\gamma}(x)$ in (2.9) has the same original form as (1.1) with η replaced by γ.

PROOF: The survival function, $\bar{G}_X(x)$, is, from (2.8)

$$\begin{aligned} \bar{G}_X(x) &= A[-u(x; \boldsymbol{\alpha}); \boldsymbol{\delta}]\exp[-\gamma u(x; \alpha)] \\ &= A[-u(x; \boldsymbol{\alpha}); \boldsymbol{\delta}]\bar{F}_{X;\gamma}(x), \end{aligned} \tag{2.10}$$

and so

$$\begin{aligned} G_X(x) &= 1 - A[-u(x; \boldsymbol{\alpha}); \boldsymbol{\delta}]\bar{F}_{X;\gamma}(x) \\ &= 1 - A[-u(x; \alpha); \boldsymbol{\delta}][1 - F_{X;\gamma}(x)]. \end{aligned}$$

It is also easy to see that the following result is true.

COROLLARY 2: If the mgf of the mixing distribution is of the form

$$M_\eta(s) = B(s; \boldsymbol{\delta})[1 - e^{\gamma s}], \qquad \gamma > 0 \tag{2.11}$$

then the compound cdf of X, $G_X(x)$, is

$$G_X(x) = 1 - B[-u(x; \boldsymbol{\alpha}); \boldsymbol{\delta}] \cdot F_{X;\gamma}(x). \tag{2.12}$$

3. COMPOUND DISTRIBUTIONS OF X WITH MULTIDIMENSIONAL MIXING DISTRIBUTIONS

Theorem 1 can be extended to the multiparameter case (1.3).

THEOREM 2: For a parental distribution, $F_X(x)$, of k-parameter exponential type defined in (1.3), when $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_k)$ are the mixing parameters with joint distribution $P_{\boldsymbol{\eta}}(y; \gamma)$ for which the mgf $M_{\boldsymbol{\eta}}(s_1, \ldots, s_k)$, exists, the compound distribution of X, $G_X(x)$, is of the form

$$G_X(x) = 1 - M_{\boldsymbol{\eta}}[-u_1(x; \boldsymbol{\alpha}_1), \ldots, -u_k(x; \boldsymbol{\alpha}_k)] \tag{3.1}$$

Proof is similar to that for Theorem 1.

We now consider some special cases of the joint mixing distributions.

First we notice, that (1.3) can be written as

$$F_{X;\eta}(x)=\begin{cases}0 & x\leq x_0\\ 1-\prod_{i=1}^{k}[1-F_{X;\eta_i}(x)], & x>x_0\end{cases} \tag{3.2}$$

where

$$F_{X;\eta_i}(x)=1-\exp[-\eta_i u_i(x;\alpha_i)], \qquad i=1,\ldots,k \tag{3.3}$$

is of simple exponential type. In case of a lifetime distribution of a certain system with k components. the parameters $\eta_1,\ldots,\eta_k$ may be associated with various properties (e.g., resistance, strength, elasticity, etc. of the material). Keeping $\eta_1,\ldots,\eta_k$ fixed, each $F_{X;\eta_i}(x)$ in (3.3) may represent the lifetime distribution of the ith component; thus $F_{X;\underset{\eta}{\wedge}}(x)$ in (3.2) would correspond to the minimum lifetime distribution for the system, or it represents distribution of first failure of the system.

In life testing, the assumption that $\eta_1,\ldots,\eta_k$ are fixed implies that the units in the test sample were obtained from a homogeneous population. This may not always be true, in particular when an industrial process undergoes some seasonal fluctuations or the units are biological organisms for which initial conditions (e.g., genetic makeup) cannot be controlled. We may then assume that $\eta_1,\ldots,\eta_k$ are independent random variables with joint cdf

$$P_{\boldsymbol{\eta}}(y)=\prod_{i=1}^{k}P_{\eta_i}(y). \tag{3.4}$$

If $M_\eta(s_i)$ is the mgf of η_i, then the joint mgf, $M_{\boldsymbol{\eta}}(s_1,\ldots,s_k)$, is

$$M_{\boldsymbol{\eta}}(s_1,\ldots,s_k)=\prod_{i=1}^{k}M_{\eta_i}(s_i). \tag{3.5}$$

We have the following theorem.

THEOREM 3: If $F_{X;\boldsymbol{\eta}}(x)$ is as defined in (3.2) and the mixing distribution, $P_{\boldsymbol{\eta}}(y)$ is as defined in (3.4), then the compound distribution of X, $G_X(x)$, is of the form

$$G_X(x)=1-\prod_{i=1}^{k}\{1-M_{\eta_i}[-u_i(x;\boldsymbol{\alpha}_i)]\}. \tag{3.6}$$

Here

$$1-M_{\eta_i}[-u_i(x;\alpha_i)]=G_i(x) \tag{3.7}$$

is the compound distribution

$$F_{X;\eta_i}(x)\underset{\eta_i}{\wedge}P_{\eta_i}(y).$$

Proof of Theorem 3 is straightforward. Extensions of Corollaries 1 and 2 are also straightforward.

4. ILLUSTRATIONS WITH PARENTAL WEIBULL DISTRIBUTION

To illustrate the techniques, we use some examples in which the parental distribution is Weibull. Most of the resulting compound distributions are not new and have been derived by other authors.

EXAMPLE 1: Weibull $(\eta) \underset{\eta}{\wedge}$ Gamma. The results in this example have already been obtained by Dubey (1968). The results are used to demonstrate the use of the moment generating function. Let

$$\bar{F}_{X;\eta}(x) = \exp\left[-\eta(x-\xi)^{\alpha}\right] \tag{4.1}$$

be the Weibull survival function. Suppose that η has gamma distribution with pdf

$$p_{\eta}(y) = \frac{\beta^{\delta}}{\Gamma(\delta)}(y-\gamma)^{\delta-1}e^{-(y-\gamma)/\beta}, \qquad y > \gamma. \tag{4.2}$$

The mgf of gamma distribution (4.2) is

$$M_{\eta}(s) = \frac{\beta^{\delta}e^{\gamma s}}{(\beta-s)^{\delta}} \tag{4.3}$$

The compound survival function, $G_X(x)$, is obtained from (2.10). We have $u(x;\ \xi,\ \alpha) = (x-\xi)^{\alpha}$; $A(s;\ \beta) = \dfrac{\beta^{\delta}}{(\beta-s)^{\delta}}$. Put $s = -(x-\xi)^{\alpha}$. Thus, from (2.10)

$$\bar{G}_X(x) = \frac{\beta^{\delta}}{[\beta+(x-\xi)^{\alpha}]^{\delta}} \times \exp[-\gamma(x-\xi)^{\alpha}], \qquad x > \xi. \tag{4.4}$$

Note that $\bar{G}_X(x)$ is the product of two survival functions. The first factor

$$\frac{\beta^{\delta}}{[\beta+(x-\xi)^{\alpha}]^{\delta}} \qquad x > \xi, \qquad \alpha > 0, \qquad \beta > 0, \qquad \delta > 0 \tag{4.5}$$

is the *Burr Type XII* survival function, [1] as was pointed out by Dubey (1968). The second factor

$$\exp[-\gamma(x-\xi)^{\alpha}], \qquad x > \xi, \qquad \alpha > 0, \qquad \gamma > 0 \tag{4.6}$$

is again a Weibull survival function with parameter γ instead of original η.

For $\gamma = 0$, the second factor in (4.4) is equal to 1, and it represents a Burr Type XII distribution (Dubey [1968]).

For $\alpha = 1$, the parental distribution (4.1) is exponential truncated from below at $x = \xi$, and (4.4) takes the form

$$\bar{G}_X(x) = \frac{\beta^{\delta}}{[(\beta-\xi)+x]^{\delta}}e^{-\gamma(x-\xi)}, \qquad x > \xi. \tag{4.7}$$

EXAMPLE 2: Distribution of the least value of k independent Weibull variables compounded by k independent exponential distributions.

Let

$$\bar{F}_{X;\eta}(x) = \exp\left[-\sum_{i=1}^{n} \eta_i (x-\xi)^{\alpha}\right] \tag{4.8}$$

for $x > \xi$, $\alpha > 0$, $\eta_i > 0$, $i = 1, 2, \ldots, k$.

Let the pdf of η_i be

$$p_{\eta_i}(y) = \beta e^{-\beta(y-\gamma_i)}, \qquad y > \gamma_i, \qquad \beta > 0 \tag{4.9}$$

for $i = 1, 2, \ldots, k$.

Then the compound survival function of (4.8) by (4.9) is, from Theorem 2

$$\bar{G}_X(x) = \frac{\beta^k}{[\beta - (x-\xi)^{\alpha}]^k} \exp\left[-\sum_{i=1}^{k} \gamma_i (x-\xi)^{\alpha}\right] \qquad x > \xi. \tag{4.10}$$

Notice that putting $k = \delta$ and $\sum_{i=1}^{k} \gamma_i = \gamma$, (4.10) becomes identical with (4.4). This is not surprising, since $\eta = \sum_{i=1}^{k} \eta_i$ is the sum of exponential variables with the same β, and so it has a gamma distribution. The truncation points, $\gamma_1, \ldots, \gamma_k$, contribute only to the second term in (4.10).

EXAMPLE 3: Weibull $(\eta) \underset{\eta}{\wedge}$ Uniform. This distribution was considered by Harris and Singpurwalla (1968).

Let $F_{X;\eta}(x)$ be a parental distribution, and suppose that pdf of η is uniform with pdf

$$p_{\eta}(y) = \begin{cases} 1/c & \text{for } a < y < a + c, \\ 0 & \text{elsewhere.} \end{cases} \tag{4.11}$$

The mgf of (4.11) can be written in the following form

$$M_{\eta}(s) = \frac{1}{cs} e^{as} (e^{cs} - 1). \tag{4.12}$$

Then from Corollaries 1 and 2, we have

$$\bar{G}_X(x) = \frac{1}{c \cdot u(x, \alpha)} \bar{F}_{X,a}(x) \cdot F_{X,c}(x). \tag{4.13}$$

When $F_{X,\eta}(x)$ is a Weibull distribution of form (4.1) we have, from (4.13)

$$\bar{G}_X(x) = \frac{1}{c(x-\xi)^\alpha} \exp\,[-a(x-\xi)^\alpha]\,\{1-\exp\,[-c(x-\xi)^\alpha]\}. \tag{4.14}$$

An alternative form of (4.14) is

$$\bar{G}_X(x) = \frac{2\exp\,[-(a+c/2)\,(x-\xi)^\alpha]}{c(x-\xi)^\alpha} \sinh\,[\tfrac{c}{2}(x-\xi)^\alpha]. \tag{4.15}$$

Other mixing distributions which can be considered but exponential, gamma, and uniform distributions are those which are most likely to be applicable to real situations.

REFERENCES

[1] Burr, I. W. "Cumulative Frequency Functions," Ann. Math. Statist., *13*, 215–232 (1942).

[2] Dubey, S. D. "A Compound Weibull Distribution," Nav. Res. Log. Quart., *15*, 179–188 (1968).

[3] Harris, C. M. and Singpurwalla, N. D. "Life Distributions Derived from Stochastic Hazard Rates," IEEE Transactions on Reliability, *R–17*, pp. 70–79 (1968).

THE NORMAL APPROXIMATION TO THE MULTINOMIAL WITH AN INCREASING NUMBER OF CLASSES*

Lionel Weiss

Cornell University

ABSTRACT

For a fixed number of classes and the number of trials increasing, the approach of the multinomial cumulative distribution function to a normal cumulative distribution function is familiar. In this paper we allow the number of classes to increase as the number of trials increases, and show that under certain circumstances the probabilities assigned to arbitrary regions by the multinomial distribution are all close to the probabilities assigned by the distribution of "rounded off" normal random variables. As the number of trials increases, the amount rounded off approaches zero. The result can be used to study the asymptotic distribution of functions of multinomial random variables.

1. NOTATION AND ASSUMPTIONS

For each n, $X_1(n), X_2(n), \ldots, X_{k(n)}(n)$ have a joint multinomial distribution, with parameters $n, p_1(n), \ldots, p_{k(n)}(n)$ where $p_i(n) > 0$ for $i = 1, \ldots, k(n)$,

$$\sum_{i=1}^{k(n)} p_i(n) = 1, \ \sum_{i=1}^{k(n)} X_i(n) = n.$$

We make the following assumptions:

(1.1) $\min_{1 \leq i \leq k(n)} [1 - p_i(n)] > \Delta$ for some $\Delta > 0$;

(1.2) $\min_{1 \leq i \leq k(n)} np_i(n)$ approaches infinity as n increases;

(1.3) $\sum_{i=1}^{k(n)} [np_i(n)]^{-1/2}$ approaches zero as n increases;

(1.4) $k(n) [np_{k(n)}(n)]^{-1/2}$ approaches zero as n increases.

Note that we do not require that $k(n)$ approach infinity as n increases, but our assumptions allow this to occur.

*This research was supported by the National Science Foundation under Grants GP 31430X2 and MPS 74-24270.

Define $Y_i(n)$ as $[np_i(n)]^{-1/2}[X_i(n)-np_i(n)]$ for $i=1, \ldots, k(n)$, so that

$$\sum_{i=1}^{k(n)} \sqrt{p_i(n)}\, Y_i(n)=0.$$

For typographical simplicity, from now on we write Y_i instead of $Y_i(n)$, p_i instead of $p_i(n)$, X_i instead of $X_i(n)$, and also do not explicitly exhibit the dependence on n of certain other quantities to be defined below.

Denote $P[Y_i=y_i;\ i=1, \ldots, k(n)-1]$ by $h_n(y_1, \ldots, y_{k(n)-1})$. $h_n(y_1, \ldots, y_{k(n)-1})$ is given as follows. If $y_1, \ldots, y_{k(n)-1}$ are such that $np_i+\sqrt{np_i}\, y_i$ is a non-negative integer for $i=1, \ldots, k(n)-1$, and

$$\sum_{i=1}^{k(n)-1} [np_i+\sqrt{np_i}\, y_i] \leqq n,$$

then

$$\log h_n(y_1, \ldots, y_{k(n)-1}) = \log n! + \sum_{i=1}^{k(n)} [np_i+\sqrt{np_i}\, y_i] \log p_i - \sum_{i=1}^{k(n)} \log\{[np_i+\sqrt{np_i}\, y_i]!\}$$

where $y_{k(n)}$ is given by the identity

$$\sum_{i=1}^{k(n)} \sqrt{p_i}\, y_i=0.$$

For other values of $(y_1, \ldots, y_{k(n)-1})$, $h_n(y_1, \ldots, y_{k(n)-1})=0$.

Now suppose $Z_1, \ldots, Z_{k(n)-1}$ have the following joint normal probability density function:

$$\left(\frac{1}{2\pi}\right)^{[k(n)-1]/2} [p_{k(n)}]^{-1/2} \exp\left\{-\frac{1}{2}\sum_{i=1}^{k(n)} z_i^2\right\}$$

where $z_{k(n)}$ is given by the identity

$$\sum_{i=1}^{k(n)} \sqrt{p_i}\, z_i=0.$$

Define the random variables $\bar{Z}_1, \ldots, \bar{Z}_{k(n)}$ as the following functions of $Z_1, \ldots, Z_{k(n)-1}$. For $i=1, \ldots, k(n)-1$, $\bar{Z}_i$ is the closest value to Z_i which makes $np_i+\sqrt{np_i}\bar{Z}_i$ an integer (positive, negative, or zero). If there are two possible values for $\bar{Z}_i$, use the smaller one. $\bar{Z}_{k(n)}$ is given by the identity

$$\sum_{i=1}^{k(n)} \sqrt{p_i}\, \bar{Z}_i=0.$$

Denote $P[\bar{Z}_i=\bar{z}_i;\ i=1,\ .\ .\ .,\ k(n)-1]$ by $g_n(\bar{z}_1,\ .\ .\ .,\ \bar{z}_{k(n)-1})$. $g_n(\bar{z}_1,\ .\ .\ .,\ \bar{z}_{k(n)-1})$ is given as follows. If $\bar{z}_1,\ .\ .\ .,\ \bar{z}_{k(n)-1}$ are such that $np_i+\sqrt{np_i}\,\bar{z}_i$ is an integer for $i=1,\ .\ .\ .,\ k(n)-1$, then $g_n(\bar{z}_1,\ .\ .\ .,\ \bar{z}_{k(n)-1})$ is equal to

$$\left(\frac{1}{2\pi}\right)^{[k(n)-1]/2}[p_{k(n)}]^{-1/2}\int_{R_n(\bar{z}_1,\ .\ .\ .,\ \bar{z}_{k(n)-1})}\cdots\int\exp\left\{-\frac{1}{2}\sum_{i=1}^{k(n)}z_i^2\right\}dz_1\ .\ .\ .\ dz_{k(n)-1}$$

where the region $R_n(\bar{z}_1,\ .\ .\ .,\ \bar{z}_{k(n)-1})$ is the set of points $(z_1,\ .\ .\ .,\ z_{k(n)-1})$ such that

$$\bar{z}_i-\frac{1}{2\sqrt{np_i}}\leqq z_i\leqq\bar{z}_i+\frac{1}{2\sqrt{np_i}}\qquad\text{for } i=1,\ .\ .\ .,\ k(n)-1,$$

and in the integrand, $z_{k(n)}$ is given by the identity

$$\sum_{i=1}^{k(n)}\sqrt{p_i}\,z_i=0.$$

$g_n(\bar{z}_1,\ .\ .\ .,\ \bar{z}_{k(n)-1})$ is zero for other values of $\bar{z}_1,\ .\ .\ .,\ \bar{z}_{k(n)-1}$.

For any event E, $\bar{E}$ denotes its negation. $\Phi(x)$ denotes the standard normal cumulative distribution function. Below we will use the following elementary inequalities:

(1.5) For any events $E_1,\ .\ .\ .,\ E_m$, $P(E_1\cap\ .\ .\ .\ \cap E_m)\geqq 1-\sum_{j=1}^{m}P(\bar{E}_j)$.

(1.6) For any $x>0$, $1-\Phi(x)<\dfrac{1}{x}$.

2. THE ASYMPTOTIC EQUIVALENCE OF h_n and g_n.

For any measurable region S_n in $(k(n)-1)$-dimenstional space, let $P_{h_n}(S_n)$ denote the probability assigned to S_n by h_n, and let $P_{g_n}(S_n)$ denote the probability assigned to S_n by g_n. We will prove the following:

THEOREM: $\lim_{n\to\infty}|P_{h_n}(S_n)-P_{g_n}(S_n)|=0$, for any sequence $\{S_n\}$, where S_n is an arbitrary measurable region in $(k(n)-1)$-dimensional space.

Before proving the theorem, we note the following application of the result. Suppose $H_n(y_1,\ .\ .\ .,\ y_{k(n)-1})$ is a function of $k(n)-1$ variables. Under our assumptions, the asymptotic distribution of $H_n(Y_1,\ .\ .\ .,\ Y_{k(n)-1})$ is exactly the same as the asymptotic distribution of $H_n(\bar{Z}_1,\ .\ .\ .,\ \bar{Z}_{k(n)-1})$. Since we can write

$$\bar{Z}_i=Z_i+\frac{\theta_i}{2\sqrt{np_i}},\qquad\text{where } |\theta_i|\leqq 1,$$

for a wide variety of functions H_n the asymptotic distribution of $H_n(\bar{Z}_1, \ldots, \bar{Z}_{k(n)-1})$ will be the same as the asymptotic distribution of $H_n(Z_1, \ldots, Z_{k(n)-1})$. Examples are given in section 3.

PROOF OF THEOREM. The theorem will be proved if we can show that

$$\frac{h_n(\bar{Z}_1, \ldots, \bar{Z}_{k(n)-1})}{g_n(\bar{Z}_1, \ldots, \bar{Z}_{k(n)-1})}$$

converges stochastically to one as n increases [2], which we proceed to do.

First we show that

$$\lim_{n\to\infty} P\left[\frac{h_n(\bar{Z}_1, \ldots, \bar{Z}_{k(n)-1})}{g_n(\bar{Z}_1, \ldots, \bar{Z}_{k(n)-1})} > 0\right] = 1.$$

It is easily varified that $E\{Z_i\}=0$, $\sigma^2_{Z_i}=1-p_i$, for $i=1, \ldots, k(n)$, cov $(Z_i, Z_j) = -\sqrt{p_i p_j}$ for $i \neq j$, $i, j=1, \ldots, k(n)$, where $Z_{k(n)}$ is defined by the identity

$$\sum_{i=1}^{k(n)} \sqrt{p_i}\, Z_i = 0.$$

Define α_i as the event $\{-1/2 < np_i + \sqrt{np_i}\, Z_i\}$ for $i=1, \ldots, k(n)$. The event

$$\left\{\frac{h_n(\bar{Z}_1, \ldots, \bar{Z}_{k(n)-1})}{g_n(\bar{Z}_1, \ldots, \bar{Z}_{k(n)-1})} > 0\right\}$$

is the same as $\{\bigcap_{i=1}^{k(n)} \alpha_i\}$, and therefore, using (1.5),

$$P\left[\frac{h_n(\bar{Z}_1, \ldots, \bar{Z}_{k(n)-1})}{g_n(\bar{Z}_1, \ldots, \bar{Z}_{k(n)-1})} > 0\right] \geqq 1 - \sum_{i=1}^{k(n)} P(\bar{\alpha}_i).$$

$$P(\bar{\alpha}_i) = \Phi\left(\frac{-1/2 - np_i}{\sqrt{np_i(1-p_i)}}\right) < \left(\frac{1/2 + np_i}{\sqrt{np_i(1-p_i)}}\right)^{-1},$$

using (1.6). Using assumptions (1.2) and (1.3), we find

$$\sum_{i=1}^{k(n)} \left(\frac{1/2 + np_i}{\sqrt{np_i(1-p_i)}}\right)^{-1}$$

approaches zero as n increases, and thus $P(\bigcap_{i=1}^{k(n)} \alpha_i)$ approaches one as n increases, or

$$\lim_{n\to\infty} \mathrm{P}\left[\frac{h_n(\bar{Z}_1, \ldots, \bar{Z}_{k(n)-1})}{g_n(\bar{Z}_1, \ldots, \bar{Z}_{k(n)-1})} > 0\right] = 1.$$

From the result of the preceding paragraph, with probability approaching one as n increases,

$$\log \frac{h_n(\bar{Z}_1, \ldots, \bar{Z}_{k(n)-1})}{g_n(\bar{Z}_1, \ldots, \bar{Z}_{k(n)-1})}$$

is equal to the sum of the following expressions:

$$\log n! \tag{2.1}$$

$$\sum_{i=1}^{k(n)} [np_i + \sqrt{np_i}\,\bar{Z}_i] \log p_i \tag{2.2}$$

$$-\sum_{i=1}^{k(n)} \log \{[np_i + \sqrt{np_i}\,\bar{Z}_i]!\} \tag{2.3}$$

$$\left(\frac{k(n)-1}{2}\right) \log 2\pi \tag{2.4}$$

$$1/2 \log p_{k(n)} \tag{2.5}$$

$$-\log \underset{R_n(\bar{Z}_1, \ldots, \bar{Z}_{k(n)-1})}{\int \cdots \int} \exp\left\{-\frac{1}{2}\sum_{i=1}^{k(n)} z_i^2\right\} dz_1 \ldots dz_{k(n)-1}. \tag{2.6}$$

Stirling's formula states that for any positive integer m,

$$\log m! = \frac{1}{2}\log 2\pi + \left(m + \frac{1}{2}\right)\log m - m + \frac{\omega(m)}{m},$$

where $|\omega(m)| < 1$. Applying the formula to (2.1), we write (2.1) as

$$\frac{1}{2}\log 2\pi + \left(n + \frac{1}{2}\right)\log n - n + \frac{\omega(n)}{n}. \tag{2.1$'$}$$

We apply Stirling's formula to (2.3), and write $\log\,(np_i + \sqrt{np_i}\,\bar{Z}_i)$ as

$$\log n + \log p_i + \log\left(1 + \frac{\bar{Z}_i}{\sqrt{np_i}}\right).$$

For any $c > 0$, define the event $\beta_i(c)$ as

$$\left\{ \frac{|\bar{Z}_i|}{\sqrt{np_i}} \leqq c \right\}, \qquad \text{for } i=1, \ldots, k(n).$$

The event $\beta_i(c)$ is implied by the event $\gamma_i(c)$ defined as $\{-cnp_i+1/2 \leqq \sqrt{np_i}\, Z_i \leqq cnp_i - 1/2\}$, and

$$P[\bar{\gamma}_i(c)] = 1-\Phi\left(c\sqrt{\frac{np_i}{1-p_i}} - \frac{1}{2\sqrt{np_i(1-p_i)}} \right) + \Phi\left(-c\sqrt{\frac{np_i}{1-p_i}} + \frac{1}{2\sqrt{np_i(1-p_i)}} \right)$$
$$= 2\left\{ 1-\Phi\left(c\sqrt{\frac{np_i}{1-p_i}} - \frac{1}{2\sqrt{np_i(1-p_i)}} \right) \right\} \leqq 2\left(c\sqrt{\frac{np_i}{1-p_i}} - \frac{1}{2\sqrt{np_i(1-p_i)}} \right)^{-1},$$

and it follows from assumptions (1.1)–(1.4) that

$$\sum_{i=1}^{k(n)} P[\bar{\gamma}_i(c)]$$

approaches zero as n increases, which implies that

$$\sum_{i=1}^{k(n)} P[\bar{\beta}_i(c)]$$

approaches zero as n increases, which in turn implies that

$$P[\bigcap_{i=1}^{k(n)} \beta_i(c)]$$

approaches one as n increases. This last fact implies that with probability approaching one as n increases, we can write

$$\log\left(1+\frac{\bar{Z}_i}{\sqrt{np_i}}\right) \qquad \text{as} \; \frac{\bar{Z}_i}{\sqrt{np_i}} - \frac{1}{2}\frac{\bar{Z}_i^2}{np_i} + \frac{1}{3}\left(\frac{\bar{Z}_i}{\sqrt{np_i}}\right)^3 \left(1+\frac{\theta_i \bar{Z}_i}{\sqrt{np_i}}\right)^{-3}$$

for $i=1, \ldots, k(n)$, where $|\theta_i| \leqq 1$ for $i=1, \ldots, k(n)$. Applying this expansion to (2.3), we find that with probability approaching one as n increases, we can write (2.3) as

$$(2.3)' \qquad -\frac{k(n)}{2}\log 2\pi + n - n\log n - \frac{k(n)}{2}\log n$$
$$-\sum_{i=1}^{k(n)} (np_i + \sqrt{np_i}\,\bar{Z}_i)\log p_i - \frac{1}{2}\sum_{i=1}^{k(n)} \log p_i - \frac{1}{2}\sum_{i=1}^{k(n)} \bar{Z}_i^2 + \Delta(n)$$

where $\Delta(n)$ is equal to

$$-\frac{1}{2}\sum_{i=1}^{k(n)}\left\{\frac{\bar{Z}_i}{\sqrt{np_i}}-\frac{1}{2}\frac{\bar{Z}_i^2}{np_i}+\frac{1}{3}\left(\frac{\bar{Z}_i}{\sqrt{np_i}}\right)^3\left(1+\frac{\theta_i\bar{Z}_i}{\sqrt{np_i}}\right)^{-3}\right\}$$

$$-\frac{1}{3}\sum_{i=1}^{k(n)}(np_i+\sqrt{np_i}\,\bar{Z}_i)\left(\frac{\bar{Z}_i}{\sqrt{np_i}}\right)^3\left(1+\frac{\theta_i\bar{Z}_i}{\sqrt{np_i}}\right)^{-3}$$

$$+\frac{1}{2}\sum_{i=1}^{k(n)}\frac{\bar{Z}_i^3}{\sqrt{np_i}}-\sum_{i=1}^{k(n)}\frac{\omega(np_i+\sqrt{np_i}\,\bar{Z}_i)}{np_i+\sqrt{np_i}\,\bar{Z}_i}$$

It is easily shown that $\Delta(n)$ converges stochastically to zero as n increases, by using assumptions (1.1)–(1.4) and the following facts:

(a) $$|\omega(np_i+\sqrt{np_i}\,\bar{Z}_i)|\leqq 1 \quad\text{and } |\theta_i|\leqq 1 \quad\text{for } i=1,\ldots,k(n).$$

(b) $$\lim_{n\to\infty} P\left[\frac{|\bar{Z}_i|}{\sqrt{np_i}}\leqq\frac{1}{2};\quad i=1,\ldots,k(n)\right]=1.$$

(c) For any positive integers r, s, the sum $S_n(r, s)$ defined as

$$\sum_{i=1}^{k(n)}\frac{|\bar{Z}_i|^r}{(\sqrt{np_i})^s}$$

converges stochastically to zero as n increases. This is true because, for all large enough n, $S_n(r, s) \leqq S_n(r, 1)$ for any s greater than one, $E\{|\bar{Z}_i|^r\}\leqq A_r<\infty$ for some finite A_r independent of n, and therefore

$$E\{S_n(r, 1)\}\leqq A_r\sum_{i=1}^{k(n)}\frac{1}{\sqrt{np_i}},$$

and this last expression approaches zero as n increases, by assumption (1.3).

Next we investigate (2.6). If $(z_1, \ldots, z_{k(n)-1})$ is any point in $R_n(\bar{Z}_1, \ldots, \bar{Z}_{k(n)-1})$, we have

$$z_i=\bar{Z}_i+\frac{\bar{\theta}_i}{2\sqrt{np_i}}\qquad\text{where } |\bar{\theta}_i|\leqq 1,\qquad\text{for } i=1,\ldots,k(n)-1.$$

Since

$$z_{k(n)}=-\frac{1}{\sqrt{p_{k(n)}}}\sum_{i=1}^{k(n)-1}\sqrt{p_i}\,z_i\qquad\text{and } \bar{Z}_{k(n)}=-\frac{1}{\sqrt{p_{k(n)}}}\sum_{i=1}^{k(n)-1}\sqrt{p_i}\,\bar{Z}_i,$$

we have

$$z_{k(n)}=\bar{Z}_{k(n)}-\frac{1}{2\sqrt{np_{k(n)}}}\sum_{i=1}^{k(n)-1}\bar{\theta}_i.$$

Then we can write

$$\sum_{i=1}^{k(n)} z_i^2 = \sum_{i=1}^{k(n)} \bar{Z}_i^2 + \Delta_n(\bar{\theta}, \bar{Z}),$$

where $\Delta_n(\bar{\theta}, \bar{Z})$ is equal to

$$\sum_{i=1}^{k(n)-1} \frac{\bar{\theta}_i \bar{Z}_i}{\sqrt{np_i}} + \sum_{i=1}^{k(n)-1} \frac{\bar{\theta}_i^2}{4np_i} - \frac{1}{\sqrt{np_{k(n)}}} \bar{Z}_{k(n)} \sum_{i=1}^{k(n)-1} \bar{\theta}_i + \frac{1}{4np_{k(n)}} \left(\sum_{i=1}^{k(n)-1} \bar{\theta}_i \right)^2.$$

$| \Delta_n(\bar{\theta}, \bar{Z}) | < \epsilon_n(\bar{Z}_1, \ldots, \bar{Z}_{k(n)-1})$ where $\epsilon_n(\bar{Z}_1, \ldots, \bar{Z}_{k(n)-1})$ is defined as

$$\sum_{i=1}^{k(n)-1} \frac{|\bar{Z}_i|}{\sqrt{np_i}} + \sum_{i=1}^{k(n)-1} \frac{1}{4np_i} + \frac{k(n)\,|\bar{Z}_{k(n)}|}{\sqrt{np_k(n)}} + \left(\frac{k(n)}{2\sqrt{np_k(n)}} \right)^2.$$

It is easily shown that $\epsilon_n(\bar{Z}_1, \ldots, \bar{Z}_{k(n)-1})$ converges stochastically to zero as n increases, using facts developed above and the fact that $|\bar{Z}_{k(n)}|$ is bounded with probability approaching one as n increases. Using the law of the mean for integrals, we can write (2.6) as

$$(2.6)' \qquad -\log \left\{ \left(\prod_{i=1}^{k(n)-1} \frac{1}{\sqrt{np_i}} \right) e^{-\frac{1}{2}\sum_{i=1}^{k(n)} \bar{Z}_i^2 - \frac{1}{2}\hat{\theta}\epsilon_n(\bar{Z}_1, \ldots, \bar{Z}_{k(n)-1})} \right\}$$

where $|\hat{\theta}| \leqq 1$.

Summing (2.1)′, (2.2), (2.3)′, (2.4), (2.5), and (2.6)′, we see that

$$\log \frac{h_n(\bar{Z}_1, \ldots, \bar{Z}_{k(n)-1})}{g_n(\bar{Z}_1, \ldots, \bar{Z}_{k(n)-1})}$$

converges stochastically to zero as n increases, completing the proof of the theorem.

3. APPLICATIONS

For each n, suppose $H_n(Y_1, \ldots, Y_{k(n)-1})$ is a quadratic in $Y_1, \ldots, Y_{k(n)-1}$. We will investigate the asymptotic distribution of $H_n(Y_1, \ldots, Y_{k(n)-1})$, under certain conditions.

Suppose M_n is a $(k(n)-1)$ by $(k(n)-1)$ orthogonal matrix, with the element in row i and column j denoted by $m_n(i,j)$, and suppose

$$m_n(1,j) = \frac{\sqrt{p_j}}{\sqrt{1-p_{k(n)}}} \qquad \text{for } j=1, \ldots, k(n)-1.$$

Let Z denote the row vector $(Z_1, \ldots, Z_{k(n)-1})$, and define the row vector $W = (W_1, \ldots, W_{k(n)-1})$ by the equation $W' = M_n Z'$. Define the row vector $T = (T_1, \ldots, T_{k(n)-1})$ as

$$\left(\frac{1}{\sqrt{p_{k(n)}}} W_1, W_2, \ldots, W_{k(n)-1}\right).$$

Then $(T_1, \ldots, T_{k(n)-1})$ are independent, standard normal random variables. Define the vectors $\bar{W}$, $\bar{T}$ as the same functions of $\bar{Z}$ as W, T are of Z. We have

$$W_i - \bar{W}_i = \sum_{j=1}^{k(n)-1} m_n(i, j)(Z_j - \bar{Z}_j),$$

and therefore

$$|W_i - \bar{W}_i| \leqq \sum_{j=1}^{k(n)-1} |m_n(i, j)| \quad |Z_j - \bar{Z}_j|$$

$$\leqq \sqrt{\sum_{j=1}^{k(n)-1} m_n^2(i, j)} \sqrt{\sum_{j=1}^{k(n)-1} (Z_j - \bar{Z}_j)^2}$$

$$= \sqrt{\sum_{j=1}^{k(n)-1} \frac{\theta_j^2}{4np_j}}$$

where $|\theta_j| \leqq 1$ for all j. Thus

$$|W_i - \bar{W}_i| \leqq \frac{1}{2} \sqrt{\sum_{j=1}^{k(n)-1} \frac{1}{np_j}} = \epsilon_n,$$

say. By assumptions (1.2) and (1.3), ϵ_n approaches zero as n increases. Finally, we have

$$|T_j - \bar{T}_j| \leqq \frac{\epsilon_n}{\sqrt{p_{k(n)}}} \qquad \text{for } j = 1, \ldots, k(n) - 1.$$

Since $H_n(Z_1, \ldots, Z_{k(n)-1})$ is a quadratic function of the components of Z, we can write $H_n(Z_1, \ldots, Z_{k(n)-1})$ as $\bar{H}_n(T_1, \ldots, T_{k(n)-1})$, where this latter function is quadratic in the components of T. Then the proper orthogonal transformation taking T into $V = (V_1, \ldots, V_{k(n)-1})$ enables us to write $\bar{H}_n(T_1, \ldots, T_{k(n)-1})$ as

$$\sum_{j=1}^{k(n)-1} \lambda_n(j) V_j^2,$$

where $\{\lambda_n(j)\}$ are the characteristic roots of the matrix of $\bar{H}_n$. Also, V consists of independent, standard normal variables. and it follows easily that

$$E\{\bar{H}_n(T_1, \ldots, T_{k(n)-1})\} = \sum_{j=1}^{k(n)-1} \lambda_n(j),$$

$$\text{Variance } \{\bar{H}_n(T_1, \ldots, T_{k(n)-1})\} = 2 \sum_{j=1}^{k(n)-1} \lambda_n^2(j).$$

Under mild conditions on the sequence of vectors $\{\lambda_n(j)\}$, the asymptotic distribution of

$$\frac{\bar{H}_n(T_1, \ldots, T_{k(n)-1}) - \sum_{j=1}^{k(n)-1} \lambda_n(j)}{\sqrt{2 \sum_{j=1}^{k(n)-1} \lambda_n^2(j)}} = Q_n,$$

say, is standard normal. Let us assume these conditions are satisfied. Define $\bar{V}$ as the same function of $\bar{T}$ as V is of T. Then $H_n(\bar{Z}_1, \ldots, \bar{Z}_{k(n)-1})$ can be written as

$$\sum_{j=1}^{k(n)-1} \lambda_n(j)\bar{V}_j^2.$$

By an argument similar to that used above to bound $|W_i - \bar{W}_i|$, we find

$$|V_i - \bar{V}_i| \leqq \left(\frac{k(n)\epsilon_n^2}{p_{k(n)}}\right)^{1/2} = \delta_n,$$

say. If δ_n and

$$\delta_n \left(\sum_{j=1}^{k(n)-1} |\lambda_n(j)|\right) \left(\sum_{j=1}^{k(n)-1} \lambda_n^2(j)\right)^{-1/2}$$

both approach zero as n increases, it is easily verified that

$$\frac{H_n(\bar{Z}_1, \ldots, \bar{Z}_{k(n)-1}) - \sum_{j=1}^{k(n)-1} \lambda_n(j)}{\sqrt{2 \sum_{j=1}^{k(n)-1} \lambda_n^2(j)}}$$

differs from Q_n by a quantity approaching zero stochastically as n increases. From the theorem proved in Section 2, it follows that the asymptotic distribution of

$$\frac{H_n(Y_1, \ldots, Y_{k(n)-1}) - \sum_{j=1}^{k(n)-1} \lambda_n(j)}{\sqrt{2 \sum_{j=1}^{k(n)-1} \lambda_n^2(j)}}$$

is standard normal.

One particular quadratic function of interest is

$$\sum_{j=1}^{k(n)} Y_j^2,$$

used in tests of fit. The asymptotic normality of this function has been proved by various authors, using assumptions under which the theorem of the present paper is not true. For a recent reference, see [1].

4. CONCLUDING REMARKS

The Theorem of Section 2 states that $\lim_{n\to\infty} |P_{h_n}(S_n) - P_{g_n}(S_n)| = 0$. It would be useful to study the rate at which $\sup_{S_n} |P_{h_n}(S_n) - P_{g_n}(S_n)|$ approaches zero as n increases, where the supremum is over all measurable regions in $(k(n)-1)$-dimensional space. That is, we would like a Berry-Esseen type bound for arbitrary measurable regions. The computations necessary to give a reasonably tight bound are lengthy.

REFERENCES

[1] Morris, C. "Central Limit Theorems for Multinomial Sums," Annals of Statistics, *3*, 165–188 (1975).

[2] Weiss, L. "Statistical Procedures Based on a Gradually Increasing Number of Order Statistics," Communications in Statistics, *2*, 95–114 (1973).

A MARKOV CHAIN VERSION OF THE SECRETARY PROBLEM

S. Christian Albright

Indiana University
Bloomington, Indiana

ABSTRACT

This paper deals with the Secretary Problem where n secretaries are interviewed sequentially and the best k must be hired. The values of the secretaries are observed as they are interviewed, but beforehand only the distributions of these values are known. Furthermore, the distributions of two successive secretaries' values are governed by a Markov chain. Optimal hiring policies for finite n and limiting optimal policies as k and n approach infinity are obtained.

1. INTRODUCTION

The Secretary Problem has been discussed in the past by many authors [1, 3–9]. The problem is that a fixed number n of secretaries are interviewed sequentially and the value of each is observed, where this value may either be a numerical value or simply a relative ranking. The case where only a relative ranking is observed has been treated by others [5, 6, 8, 9], and the necessary analysis and even the results are of a quite different nature than the case treated here, namely where the value of a secretary is a numerical value from a known probability distribution. Immediately after a secretary's value is observed, a decision must be made to either hire her or not, with the proviso that she cannot be hired later on. The objective is to fill the k available jobs, with k possibly greater than 1, with the k best secretaries. In this paper, we assume the values of the successive secretaries are random variables from r different known distributions such that the successive distributions are governed by a Markov chain with a known transition matrix. The objective is to find a sequential policy which maximizes the expected sum of the values of those k secretaries chosen for the k positions.

In the first section below we set up the problem and calculate the optimal policy for finite n. In the next section the limiting form of this policy as $n \to \infty$ is discovered. In the final section a slightly different problem is formulated and its optimal policy for finite n and its limiting optimal policy are given.

We remark at this time that the context of secretaries interviewing for jobs is retained mainly because of historical interest. The problem discussed certainly has other, probably more interesting, applications anytime a decision-maker wishes to choose the k best of n sequentially arriving objects. A possible example is where there are several assembly lines (or several production plants) which jointly must fill k special orders from the next n items produced. The lines may differ as to quality and the probability that a given line is the next to produce an item may depend on the line which produced the previous item, as confirmed by historical records. If a decision must be made on each item as soon as it is produced, then our model can be used to sequentially choose the best k in an optimal manner.

2. OPTIMAL POLICY FOR FINITE n

Let $T_1, T_2, \ldots$ be the successive states of an r — state Markov chain with transition probabilities q_{ij}. Associate with each state i a sequence of i.i.d. non-negative random variables $X_1^i, X_2^i, \ldots$ from distribution F_i. Then we assume the value of the jth secretary is a random variable X_j, where

$$X_j = \sum_{i=1}^{r} X_j^i 1_{Tj=i},$$

and 1_A is the indicator of the set A. We assume that when a secretary is interviewed, both the value of X_j and the distribution it comes from are observed. The following theorem gives the optimal policy for any k and n.

THEOREM 1: Assume there are n secretaries left to interview. Then there are numbers $0 \equiv a_{0,n}^i \leqslant a_{1,n}^i \leqslant \ldots \leqslant a_{n-1,n}^i \leqslant a_{n,n}^i \equiv \infty$, such that if there are $k \leqslant n$ jobs still vacant and the next secretary has a value x from distribution F_i, she should be hired if and only if $x \geqslant a_{n-k,n}^i$.

Furthermore, suppose that in an n secretary problem, the first secretary's value comes, from distribution F_i, a decision is made on her, and then $n-1$ secretaries and k vacancies remain. At this point the sum

$$\sum_{j=1}^{k} a_{n-j,n}^i$$

is the optimal expected reward from the k secretaries who will eventually be chosen. That is, $a_{n-k,n}^i$ may be interpreted as the incremental value from having k vacancies instead of $k-1$.

Finally, the critical values may be calculated from the recursion:

$$(1) \qquad a_{m,n+1}^i = \sum_{j=1}^{r} \left\{ \int_{a_{m-1,n}^j}^{a_{m,n}^j} xF_j(dx) + a_{m-1,n}^j F_j(a_{m-1,n}^j) + a_{m,n}^j (1 - F_j(a_{m,n}^j)) \right\} q_{ij}, \qquad 1 \leqslant i \leqslant r.$$

where the convention is that $\infty \cdot 0$ when $m=n$.

PROOF: The proof will be omitted since it is a straightforward induction proof on n. The recursion (1) follows from the interpretation of $a_{m,n+1}^i$ given above.

3. LIMITING RESULTS FOR LARGE n

This section deals with the case where $n \to \infty$ but k/n approaches a nonzero limit. In particular we will find $\lim_{n\to\infty} a_{[n\pi],n+1}^j$ and $\lim_{n\to\infty} A_j(n,\pi)/n$ for a fixed fraction $0 < \pi < 1$, where

$$A_j(n,\pi) = \sum_{m=[n\pi]+1}^{n} a_{m,n+1}^j$$

and $[n\pi]$ is the greatest integer in $n\pi$. Note that from Theorem 1, $A_j(n,\pi)$ is the optimal total expected value from n secretaries filling $n-[n\pi]$ vacancies, given that the previous secretary was from distribution j. We find the limit of this quantity divided by n by comparing the optimal policy with a sub-

optimal policy to obtain a lower bound and an impossible but better than optimal policy to obtain an upper bound.

First we define some notation. Assume the Markov chain governing the successive distributions is ergodic and let $\rho_1, \ldots, \rho_r$ be the associated steady-state probabilities. Let $l_i(n)$ be the number of X's from distribution F_i out of the first n, and let $\nu_1(i), \nu_2(i), \ldots$ denote the indices of these X's. Define

$$Z_j^i(a) = 1_{(X_j^i \geq a)} \quad \text{and } Y_j^i(a) = X_j^i 1_{(X_j^i \geq a)},$$

and then for any a and b, let

$$s_i(n) \equiv s_i(n; a, b) = \min \left(m \middle| \sum_{j=1}^{m} Z_{\nu_j(i)}^i(a) \geq [nb] \right)$$

$$t_i(n) \equiv t_i(n; a, b) = \min (l_i(n), s_i(n)).$$

Finally, for any $0 \leq \sigma_i \leq 1$ and $1 \leq i \leq r$, denote by $F_i^{-1}(\sigma_i)$ any of the numbers τ which satisfy $F_i(\tau) = \sigma_i$, and let

$$\mu_i(\sigma_i) = \int_{F_i^{-1}(\sigma_i)}^{\infty} x F_i(dx).$$

The following three lemmas will establish the lower bound mentioned above.

LEMMA 1: Let $0 \leq \sigma_i \leq 1$ be any number for which $\{\tau | F_i(\tau) = \sigma_i\}$ is nonempty, and let $s_i(n) = s_i(n; F_i^{-1}(\sigma_i), \rho_i(1-\sigma_i))$. Then

$$\lim_{n \to \infty} E t_i(n)/n = \rho_i.$$

PROOF: Since $0 \leq t_i(n)/n \leq l_i(n)/n \leq 1$, it suffices by the bounded convergence theorem to show that $t_i(n)/n \to \rho_i$ in probability as $n \to \infty$. Let $\epsilon > 0$. Then

$$P(t_i(n)/n \geq \rho_i + \epsilon) \leq P(l_i(n)/n \geq \rho_i + \epsilon) \to 0 \text{ as } n \to \infty$$

by Markov chain theory. Also

$$P(t_i(n)/n \leq \rho_i - \epsilon) \leq P(l_i(n)/n \leq \rho_i - \epsilon) + P(s_i(n)/n \leq \rho_i - \epsilon)$$

and again the first term on the right goes to 0. For the second term, we have

$$P(s_i(n)/n \leq \rho_i - \epsilon) = P\,(s_i(n) \leq [n(\rho_i - \epsilon)])$$

$$= P\left\{ \sum_{j=1}^{[n(\rho_i - \epsilon)]} Z_{\nu_j(i)}^i(F_i^{-1}(\sigma_i)) \geq [n\rho_i(1-\sigma_i)] \right\}$$

$$=P\left\{\sum_{j=1}^{[n(\rho_i-\epsilon)]} Z_j^i(F_i^{-1}(\sigma_i)) \geqslant [n\rho_i(1-\sigma_i)]\right\}$$

$$=P\left\{\sum_{j=1}^{[n(\rho_i-\epsilon)]} \frac{Z_j^i(F_i^{-1}(\sigma_i))}{[n(\rho_i-\epsilon)]} \geqslant \frac{[n\rho_i(1-\sigma_i)]}{[n(\rho_i-\epsilon)]}\right\}$$

and this term goes to 0 by the weak law of large numbers, since the Z's are i.i.d. with mean $1-\sigma_i$ and

$$\lim_{n\to\infty} \frac{[n\rho_i(1-\sigma_i)]}{[n(\rho_i-\epsilon)]} = \frac{1-\sigma_i}{1-\epsilon/\rho_i} > 1-\sigma_i.$$

LEMMA 2: Let σ_i, $s_i(n)$ and $t_i(n)$ be as in Lemma 1. Then

$$\lim_{n\to\infty} E\left\{\sum_{j=1}^{t_i(n)} Y^i_{\nu_j(i)}(F_i^{-1}(\sigma_i))\right\} \Big/ n = \rho_i\mu_i(\sigma_i).$$

PROOF: Define $A(l)$ to be the set of possible realizations of $\nu_1(i)$, $\nu_2(i)$, . . . when $l_i(n)=l$. Also, for $\nu \equiv (\nu_1, \nu_2, \ldots) \epsilon A(l)$, let $A(l, \nu) = \{l_i(n) = l;\ \nu_j(i) = \nu_j, \text{ all } j\}$. Notice that $A(l, \nu)$ is simply a subset of the probability space associated with the Markov chain. Furthermore, $A(l)$ and $A(l, \nu)$ are obviously dependent on n and i, although this dependence has been suppressed for notational convenience. Then,

$$E\left\{\sum_{j=1}^{t_i(n)} Y^i_{\nu_j(i)}(F_i^{-1}(\sigma_i))\right\} \Big/ n = \sum_l \sum_{\nu\epsilon A(l)} E\left\{\sum_{j=1}^{t_i(n)} Y^i_{\nu_j(i)}(F_i^{-1}(\sigma_i)) \,\Big|\, A(l,\nu)\right\} P(A(l,\nu)) \Big/ n.$$

On the set $A(l, \nu)$, $t_i(n)$ is a finite stopping rule for the i.i.d. $Y^i_{\nu_j}$'s, so by Wald's equation and independence the above expression is

$$\sum_l \sum_{\nu\epsilon A(l)} E(t_i(n) \mid A(l,\nu)) E\{Y_1^i(F_i^{-1}(\sigma_i))\} P(A(l,\nu)) \Big/ n = (Et_i(n)/n)\mu_i(\sigma_i) \to \rho_i\mu_i(\sigma_i) \text{ as } n\to\infty,$$

by Lemma 1.

LEMMA 3: Suppose $F_1, \ldots, F_r$ are continuous. Then for any $0<\pi<1$ there exist fractions $0 \leqslant \sigma_i \leqslant 1$, $1 \leqslant i \leqslant r$, for which $\Sigma\rho_i\sigma_i = \pi$ and $F_1^{-1}(\sigma_1) = \ldots = F_r^{-1}(\sigma_r)$. Furthermore, for these σ_i's and any $1 \leqslant j \leqslant r$,

$$\liminf_{n\to\infty} A_j(n,\pi) \Big/ n \geqslant \sum_{i=1}^{r} \rho_i\mu_i(\sigma_i).$$

PROOF: The existence of the σ_i's follows by looking at solutions of $F_1^{-1}(\sigma_1) = \ldots = F_r^{-1}(\sigma_r) = b$ and then adjusting b until $\Sigma\rho_i\sigma_i = \pi$.

Recall that we are trying to fill $n-[n\pi]$ vacancies with n secretaries. Consider the infeasible policy which hires at most $[n\rho_i(1-\sigma_i)]$ of the secretaries from distribution F_i and, with this restric-

tion, hires only those secretaries from F_i with a value $x \geq F_i^{-1}(\sigma_i)$. This policy is infeasible because the vacancies may not all be filled, but since

$$\sum_{i=1}^{r} [n\rho_i(1-\sigma_i)] \leq \sum_{i=1}^{r} n\rho_i(1-\sigma_i) = n(1-\pi) \leq n-[n\pi],$$

the policy will never hire too many secretaries. If R_n is the value obtained from this policy, then the part of ER_n/n due to the X^i's is exactly what we examined in Lemma 2, and its limit as $n \to \infty$ is given in Lemma 2.

Now compare the above policy with a feasible, nonoptimal policy which hires only those secretaries from F_i with a value $x \geq F_i^{-1}(\sigma_i)$ until there are as many secretaries left to interview as there are vacancies, if this ever occurs. Then hire all remaining secretaries. If $\bar{R}_n$ is the value obtained from this policy, then the non-negativity of the X's implies that

$$\sum_{i=1}^{r} \rho_i\mu_i(\sigma_i) = \lim_{n\to\infty} ER_n/n \leq \liminf_{n\to\infty} E\bar{R}_n/n \leq \liminf_{n\to\infty} A_j(n,\pi)/n.$$

Next we find an upper bound on the optimal expected reward.

LEMMA 4: Assume $F_1, \ldots, F_r$ are continuous and let $\sigma_1, \ldots \sigma_r$ be as in Lemma 3. Then

$$\limsup_{n\to\infty} A_j(n,\pi) \Big/ n \leq \sum_{i=1}^{r} \rho_i\mu_i(\sigma_i).$$

PROOF: Recall that X_j is the value of the jth secretary and let $X_{(1)}, \ldots, X_{(n)}$ be the order statistics associated with $X_1, \ldots, X_n$. Then since no sequential hiring policy can do better than hiring the secretaries with values $X_{([n\pi]+1)}, \ldots, X_{(n)}$, it suffices to show that

$$\limsup_{n\to\infty} E\left\{\sum_{m=[n\pi]+1}^{n} X_{(m)} - \sum_{i=1}^{r}\sum_{j=1}^{n} X_j 1_{A(i,j)}\right\} \Big/ n \leq 0 \tag{2}$$

where $A(i, j)$ is the event that X_j is from distribution F_i and $X_j \geq F_i^{-1}(\sigma_i) = F_1^{-1}(\sigma_1)$. This suffices because

$$\lim_{n\to\infty} E\left\{\sum_{i=1}^{r}\sum_{j=1}^{n} X_j 1_{A(i,j)}\right\} \Big/ n = \sum_{i=1}^{r} \rho_i\mu_i(\sigma_i).$$

Define $N_i(n) = \sum_{j=1}^{n} 1_{A(i,j)}$ and $N(n) = \sum_{i=1}^{r} N_i(n)$. Then the right-hand double sum in (2) contains $N(n)$ nonzero terms, all greater than or equal to $F_1^{-1}(\sigma_1)$, whereas the left-hand sum contains the $n-[n\pi]$ largest X_j's. Therefore the expectation in (2) may be written as

$$\left\{ \int_{(N(n)\leq n-[n\pi])} (n-[n\pi]-N(n) \text{ terms, all} \leq F_1^{-1}(\sigma_1)) \right.$$

$$\left. + \int_{(N(n)>n-[n\pi])} (N(n)-n+[n\pi] \text{ terms, all} \leq -F_1^{-1}(\sigma_1)) \right\} \Big/ n \leq F_1^{-1}(\sigma_1) E(n-[n\pi]-N(n) \Big/ n.$$

However, since $\lim_{n\to\infty} EN_i(n)/n = \rho_i(1-\sigma_i)$ and $\Sigma\rho_i(1-\sigma_i)=1-\pi$, the above expectation goes to 0. This shows that (2) holds, and the proof is complete.

We put the preceding lemmas together to establish the following theorem.

THEOREM 2: Let $F_1, \ldots, F_r$ and $\sigma_1, \ldots, \sigma_r$ be as in Lemmas 3 and 4. Then

$$\lim_{n\to\infty} A_j(n,\pi) \Big/ n = \sum_{i=1}^{r} \rho_i \mu_i(\sigma_i).$$

Corollary 1 below lists two interesting results on sums of order statistics associated with Markov chains. The proofs follow easily from Lemma 4 and Theorem 2.

COROLLARY 1: Under the same assumptions as in Theorem 2, we have

$$\lim_{n\to\infty} E \sum_{m=[n\pi]+1}^{n} X_{(m)} \Big/ n = \sum_{i=1}^{r} \rho_i \mu_i(\sigma_i)$$

and

$$\lim_{n\to\infty} E \left\{ \sum_{m=1}^{[n\pi]} X_{(m)} \right\} \Big/ n = \sum_{i=1}^{r} \rho_i(\mu_i(0)-\mu_i(\sigma_i)).$$

The following proposition perhaps gives a more intuitive reason why the σ_i's of Theorem 2 are associated with the *optimal* hiring policy. Its proof is straightforward and hence is omitted.

PROPOSITION 1: If $F_1, \ldots, F_r$ and $\sigma_1, \ldots, \sigma_r$ are as in Theorem 2, then these σ_i's also maximize $\Sigma\rho_i\mu_i(\sigma_i)$ subject to $\Sigma\,\rho_i\sigma_i=\pi$, $0\leq\sigma_i\leq 1$.

Finally, we use Theorem 2 to obtain a limit on $a^j_{[n\pi],\,n+1}$ itself. Assuming the σ_i's are as in Theorem 2, define $x(\pi)$ to be the common number $F_1^{-1}(\sigma_1)= \ldots =F_r^{-1}(\sigma_r)$. Also define $g(x)=\Sigma\rho_i F_i(x)$ and assume that g is strictly increasing in a neighborhood around $x(\pi)$. Then g^{-1} is defined in a neighborhood around π since $g(x(\pi))=\pi$, so that $x(\pi)=g^{-1}(\pi)$. We use this to obtain the following theorem.

THEOREM 3: If $F_1, \ldots, F_r$ are each absolutely continuous and g and $x(\pi)$ are defined as above, then for any $0<\pi<1$, $\lim_{n\to\infty} a^j_{[n\pi],\,n+1}=x(\pi)$.

PROOF: Since the proof is very similar to the analogous proof of Theorem 2 in [2], it will be omitted.

The above theory has two important implications for implementation. First, although the optimal policy in Theorem 1 is simple to administer if the $a^i_{m,n}$'s are known, these critical numbers may be difficult to calculate. By Theorem 3, however, there are simple approximations for the critical numbers which may be used instead, at least for large n, to achieve near optimality. Second, the feasible, non-optimal policy described in the proof of Lemma 3 also yields nearly optimal results. Furthermore,

this policy is trivial to administer and calculate since there is a single time-independent critical number for each distribution which should be used until there are as many vacancies as secretaries to be interviewed. Past this point all secretaries would be hired.

Finally, we mention a special case of the above model which is probably more realistic in many situations. This is when the next distribution is independent of the most recent one, that is, when $q_{ij}=q_j$ for each i, j. However, our results then hold by taking $r=1$ and using the distribution $F=\Sigma q_i F_i$ which is a random mixture of the given distributions.

4. ANOTHER SECRETARY PROBLEM

Suppose the above model is changed so there are r categories of jobs such that a secretary from distribution F_i can be used only for vacancies of type i. In this case it is convenient to think of r decisionmakers, each dealing with a single category of jobs, and each free to make his decisions independently of the others. Also, we should now think of the number of available type i jobs as simply a *maximal* number of openings, so that if less than this many type i secretaries arrive, there is no penalty for unfilled positions. For this problem we list results analogous to those in the previous two sections. Since the proofs of these contain no really different ideas than those in the previous sections, we omit them.

THEOREM 4: Assume there are n secretaries remaining to be seen. Then there are numbers $0\equiv a^i_{0,n}\leq a^i_{1,n}\leq \ldots \leq a^i_{n-1,n}\leq a^i_{n,n}\equiv\infty$, such that if there are $k\leq n$ vacancies of type i and the next secretary has a value x from distribution F_i, then she should be hired if and only if $x\geq a^i_{n-k,n}$.

Furthermore, if $n-1$ more secretaries remain to be seen, there are k vacancies of type i, and the previous secretary was from distribution F_i, then the sum

$$\sum_{j=1}^{k} a^i_{n-j,n}$$

is the optimal expected reward from hiring the type i secretaries. That is, $a^i_{n-k,n}$ is the incremental value from having k openings of type i instead of only $k-1$.

Finally, these critical numbers may be calculated from the recursions:

$$a^i_{1,n+1}=q_{ii}\left\{\int_0^{a^i_{1,n}} xF_i(dx)+a^i_{1,n}\,(1-F_i(a^i_{1,n}))\right\}$$

and

$$a^i_{j,n+1}=(1-q_{ii})a^i_{j-1,n}+q_{ii}\left\{\int_{a^i_{j-1,n}}^{a^i_{j,n}} xF_i(dx)+a^i_{j-1,n}\,F_i(a^i_{j-1,n})+a^i_{j,n}\,(1-F_i(a^i_{j,n}))\right\},$$

for $2\leq j\leq n$.

Now suppose n secretaries remain to be seen, the previous secretary was of type i, and $n-[n\pi]$ vacancies of type i remain, where $0<\pi<1$. Then the optimal total expected reward from the secretaries hired to these jobs is, by Theorem 4, $A_i(n, \pi)$. Again we obtain a limit theorem for this quantity in the interesting case where, in the limit, there are not enough type i jobs. As before, we assume the Markov chain is ergodic with stationary distribution $\{\rho_i\}$.

THEOREM 5: For $1-\rho_i<\pi<1$, let $\sigma_i=1-(1-\pi)/\rho_i$, and assume $\{\tau \mid F_i(\tau)=\sigma_i\}$ is non-empty. Then $\lim_{n\to\infty} A_i(n, \pi)/n=\rho_i\mu_i(\sigma_i)$.

THEOREM 6: Suppose F_i is absolutely continuous and π and σ_i are as in Theorem 5. Then

$$\lim_{n\to\infty} a^i_{[n\pi], n+1}=F_i^{-1}(\sigma_i).$$

Also, let $R_i=\sup\{x \mid F_i(x)=0\}$. Then for $0<\pi<1-\rho_i$,

$$\limsup_{n\to\infty} a^i_{[n\pi], n+1}\leq \underline{R}_i$$

We conclude this section by describing one possible setting for the model of this section. Suppose the different categories of vacancies refer to different levels of necessary qualifications. For example, level 1 might be relatively menial work, whereas a higher level might be a more responsible position. Also assume that the values of all secretaries come from one distribution F, but there are fixed numbers $0\equiv t_0\leq t_1\leq \ldots \leq t_{r-1}\leq t_r\equiv\infty$, such that a value x satisfying $t_{i-1}<x<t_i$ can only be considered for level i. This fits our model if we let $q_{ji}=q_i=F(t_i)-F(t_{i-1})$ and

$$F_i(x)=\begin{cases}\dfrac{F(x)-F(t_{i-1})}{F(t_i)-F(t_{i-1})} & t_{i-1}<x<t_i\\ 0 & x\leq t_{i-1}\\ 1 & x>t_i\end{cases}$$

5. CONCLUDING REMARKS

Before concluding, we make several remarks which are pertinent to the above results. In section 2 it is possible that the distributions $F_1, \ldots, F_r$ are degenerate at the points $x_1, \ldots, x_r$, respectively. Then we would actually be observing a Markov chain, where being in state i would mean observing x_i. In this case equation (1) takes the form

$$a^i_{m, n+1}=\sum_{j=1}^{r}\left\{x_j 1_{(a^j_{m-1, n}<x_j<a^j_{m, n})}+a^j_{m-1, n}1_{(x_j\leq a^j_{m-1, n})}+a^j_{m, n}1_{(x_j>a^j_{m, n})}\right\}q_{ij}, \qquad 1\leq i\leq r, \tag{1'}$$

where $1_{(x\in A)}$ is 1 if $x\in A$ and 0 otherwise.

Second, as in [1, 2, 3] any of the above results may be generalized to the case where there are n fixed quantities $p_1\leq \ldots \leq p_n$, or r sets of p's in the model of section 4, which must be assigned to the n X's, so that a p assigned to an x receives a reward px. (Our p's were all 0's and 1's.) The basic results are that the same sequences of $a^i_{m,n}$'s determine the optimal policy for *any* set of p's. For example in the model of section 2, if n p's remain and the next x satisfies $a^i_{j-1,n}\leq x<a^i_{j,n}$, then p_j should be assigned to this x. The reader is referred to [1] for more complete results along this line.

Finally, further work should be done on the case where only the values of the x's, not their distributions, are observed. In some contexts this would seem to be the most realistic approach.

However, there are also many cases where qualitative differences in the secretaries, or whatever objects are being observed, immediately identify the distributions from which they come, and this is the rationale for the approach taken here.

REFERENCES

[1] Albright, S. C., "Stochastic Sequential Assignment Problems," Technical Report No. 147 (Unpublished doctoral dissertation), Operations Research Department, Stanford University, 1972.

[2] Albright, S. C., and C. Derman, "Asymptotic Optimal Policies for the Stochastic Sequential Assignment Problem," Management Science, *19*, 46–51 (1972).

[3] Derman, C., G. J. Lieberman, and S. M. Ross, "A Sequential Stochastic Assignment Problem," Management Science, *18*, 349–355 (1972).

[4] Elfving, G., "A Persistency Problem Connected with a Point Process," Journal of Applied Probability, *4*, 77–89 (1967).

[5] Gilbert, J., and F. Mosteller, "Recognizing the Maximum of a Sequence," Journal of the American Statistical Association, *61*, 35–73 (1966).

[6] Guesin-Zade, S. M., "The Problem of Choice and the Optimal Stopping Rule for a Sequence of Independent Trials," Theory of Probability and Its Applications, *11*, 472–476 (1966).

[7] Karlin, S., Optimal Disposal of an Asset, Studies in Applied Probability and Management Science, K. J. Arrow, S. Karlin, H. Scarf, editors, Stanford University Press, Stanford, California (1962).

[8] Rasmussen, W. T., "A Generalized Secretary Problem: Choosing any of the d Best from a Sequence of Independent Values," Technical Report, Naval Electronics Laboratory Center, San Diego, California, 1974.

[9] Rasmussen, W. T., and Robbins, Herbert, "The Candidate Problem with Unknown Population Size," Technical Report, Naval Electronics Laboratory Center, San Diego, California (1974).

A NOTE ON EXPLICIT SOLUTION IN LINEAR FRACTIONAL PROGRAMMING*

A. Charnes

The University of Texas at Austin

Daniel Granot and Frieda Granot

University of British Columbia
Vancouver, B.C., Canada

ABSTRACT

In this note we analyze the fractional interval programming problem (FIP) and find, explicitly, all its optimal solutions.

Though our results are essentially the same as those in Charnes and Cooper [4], the proofs and analysis we provide here are considerably simpler.

0. INTRODUCTION

Charnes and Cooper [3] provided a complete analysis of the linear fractional programming problem in all generality, and reduced the fractional problem to at most a pair of ordinary linear programming problems. They further applied the results in [3] to obtain an explicit optimal solution to a general class of linear fractional problems — those for which the constraint set is given by $b^- \leqslant Ax \leqslant b^+$ and A is of full row rank, see [4]. These problems were termed fractional interval programming problems (FIP) see [1, 4].

In this note we analyze the full row rank (FIP) and find, explicitly, all its optimal solutions. Though our results are essentially the same as those in [4], the proofs and analysis we provide here are considerably simpler.

We transform the (FIP) problem to an equivalent problem in which all the regular cases are easily detected. Further, unlike in [4], we do not apply the Charnes and Cooper transformation from a fractional linear problem to an ordinary linear problem in order to find an explicit optimal solution to (FIP). Rather, the optimality of the solutions we generate follows from a well-known property of the fractional function.

1. THE (FIP) PROBLEM

Consider the fractional interval programming problem (FIP) of the form

$$\text{Max} \frac{\hat{c}^T x + \hat{c}_0}{\hat{d}^T x + \hat{d}_0} \tag{1}$$

*This research was partially supported by N.R.C. Grant No. A-4024 and by Office of Naval Research Contracts N00014-67-A-0126-0008 and N00014-67-A-0009 with the Center for Cybernetic Studies, The University of Texas, Austin, Texas.

subject to

(2) $$b^{-} \leqslant Ax \leqslant b^{+}$$

where A is of full row rank (see [1, 4]).

Let us assume that (FIP) is feasible, i.e.

(3) $$S = \{x \epsilon R^{n}; \qquad b^{-} \leqslant Ax \leqslant b^{+}\} \neq \emptyset.$$

ASSUMPTION 1: $\hat{d}^{T}x + \hat{d}_0 \neq$ constant on S.

ASSUMPTION 2: There does not exist a constant $\lambda \epsilon R$ such that $(\hat{c}^{T}x + \hat{c}_0) = \lambda(\hat{d}^{T}x + \hat{d}_0)$ on S.

Since A is of full row rank it has a right inverse denoted by $A^{\#}$. Substituting

(4) $$\hat{y} = Ax$$

or

(5) $$x = A^{\#}\hat{y} + (I - A^{\#}A)z, \qquad z \text{ arbitrary}$$

in (1), (2) we obtain an equivalent problem to (FIP) of the form

(6) $$\text{Max } \frac{\hat{c}^{T}[A^{\#}\hat{y} + (I - A^{\#}A)z] + \hat{c}_0}{\hat{d}^{T}[A^{\#}\hat{y} + (I - A^{\#}A)z] + \hat{d}_0}$$

subject to

(7) $$b^{-} \leqslant \hat{y} \leqslant b^{+}.$$

In order that (FIP), under assumptions 1, 2, will be bounded from above, we further assume that $\hat{c}^{T} \perp (I - A^{\#}A)$ and $\hat{d}^{T} \perp (I - A^{\#}A)$. A complete analysis when $\hat{c}^{T} \not\perp (I - A^{\#}A)$ or $\hat{d}^{T} \not\perp (I - A^{\#}A)$ is given in Charnes and Cooper [4].

Without loss of generality we can assume that

(8) $$b_i^{-} < b_i^{+} \qquad (i = 1, \ldots, m).$$

Otherwise, if $b_i^{-} > b_i^{+}$ then (FIP) is infeasible, or, if $b_i^{-} = b_i^{+}$ we substitute $\hat{y}_i = b_i^{+}$.

Denoting by

(9) $$y_i = \frac{\hat{y}_i - b_i^{-}}{b_i^{+} - b_i^{-}} \qquad (i = 1, \ldots, m)$$

and substituting in (6) and (7) results with the following equivalent problem:

(10) $$\text{Max } \left\{ \frac{c^{T}y + c_0}{d^{T}y + d_0} = V(y) \right\}$$

subject to

$$0 \leq y \leq 1 \tag{11}$$

where

$$c^T = \hat{c}^T A^\#, \qquad d^T = \hat{d}^T A^\#, \qquad c_0 = \hat{c}_0, \qquad d_0 = \hat{d}_0.$$

We can assume, without loss of generality, that $d_i \geq 0$ $(i = 1, \ldots, m)$, since otherwise, if $d_k < 0$ we substitute $\tilde{y}_k = 1 - y_k$.

Without loss of generality* the following three disjoint and exhaustive cases are describing the behavior of the denominator of $V(y)$ on $S_1 = \{y; 0 \leq y \leq 1\}$.

I. $d_0 < 0$, $d_0 + \sum_{i=1}^{m} d_i > 0$ (i.e. the denominator changes sign on S_1)

II. $d_0 = 0$ (i.e. the denominator has a unique sign but vanishes on S_1)

III. $d_0 > 0$ (i.e. the denominator does not vanish on S_1).

CASE I: Here there are the following two subcases

(a) For some

$$0 < y < 1, \qquad d_0 + \sum_{i=1}^{m} d_i y_i = 0 \qquad \text{and } c_0 + \sum_{i=1}^{m} c_i y_i > 0.$$

Then, $V(y)$ is unbounded on S_1, see Charnes and Cooper [4].

(b) For every $0 \leq y \leq 1$, such that

$$d_0 + \sum_{i=1}^{m} d_i y_i = 0$$

we have

$$c_0 + \sum_{i=1}^{m} c_i y_i = 0.$$

Charnes and Cooper [4] proved, by the aid of Farkas-Minkowski lemma, that if subcase I(b) holds

---

*Since we can always multiply the numerator and the denominator of $V(y)$ by -1 and substitute for each y_i, $i \in \{1, \ldots, m\}$ for which $d_i < 0$ a new variable $\tilde{y}_i = 1 - y_i$ case II above accounts for the situation where

$$d_0 < 0 \qquad \text{and } d_0 + \sum_{i=1}^{m} d_i = 0$$

whereas case III accounts for the situation where

$$d_0 < 0 \qquad \text{and } d_0 + \sum_{i=1}^{m} d_i < 0,$$

and therefore the three cases defined above are disjoint and exhaustive.

then $V(y)$ is constant over S_1. A different elementary proof will be given here.

It is easy to verify for I(b) that $c_i \doteq 0$ must hold whenever $d_i = 0$. Moreover, there exists a vector $y^1 = (y_1^1, \ldots, y_m^1)$, $0 < y^1 < 1$ for which

$$d_0 + \sum_{i=1}^{m} d_i y_i^1 = 0.$$

Next, let us construct the following vectors

$$y^2 = \left(y_1^1 - \epsilon, y_2^1, \ldots, y_{m-1}^1, y_m^1 + \frac{d_1}{d_m} \cdot \epsilon \right), \ldots, y^m = \left(y_1^1, \ldots, y_{m-2}^1, y_{m-1}^1 - \epsilon, y_m^1 + \frac{d_{m-1}}{d_m} \epsilon \right)$$

Clearly, for $\epsilon > 0$ and sufficiently small we have that $0 < y^i < 1$ and

$$d_0 + d^T y^i = 0 \qquad (i = 2, \ldots, m).$$

LEMMA 1: The vectors $y^1, \ldots, y^m$ constructed above are linearly independent.

PROOF: Assume the contrary. Then

$$\alpha_1 y^1 + \ldots + \alpha_m y^m = 0 \tag{12}$$

implies

$$\begin{cases} y_{i-1}^1(\alpha_1 + \ldots + \alpha_m) - \alpha_i \cdot \epsilon = 0 & (i = 2, \ldots, m) \\ y_m^1(\alpha_1 + \ldots + \alpha_m) + \epsilon(\alpha_2 \dfrac{d_1}{d_m} + \alpha_3 \dfrac{d_2}{d_m} + \ldots + \alpha_m \dfrac{d_{m-1}}{d_m}) = 0 & \end{cases} \tag{13}$$

Now, if for some $2 \leqslant i \leqslant m$, $\alpha_i = 0$, then $\alpha_j = 0$, $j = 1, \ldots, m$. Thus, if $y^1, \ldots, y^m$ are dependent we must have $\alpha_i \neq 0$ $(i = 2, \ldots, m)$. Further, since $\epsilon > 0$, $y_i^1 > 0$ $(i = 1, \ldots, m)$ we obtain

$$\text{sign}\,(\alpha_1 + \ldots + \alpha_m) = \text{sign}\,(\alpha_i) \qquad (i = 2, \ldots, m). \tag{14}$$

Now, since $d_1 > 0 (i = 1, \ldots, m)$ we conclude that the last equation in (13) cannot hold, which completes the proof.

Let y be any element in S_1 for which $d_0 + d^T y \neq 0$. Then, since y^i $(i = 1, \ldots, m)$ span R^m, y can be uniquely represented as

$$y = \sum_{i=1}^{m} \beta_i y^i. \tag{15}$$

Substituting (15) in $V(y)$ results in

(16)
$$V(y)=\frac{c_0+\sum_{i=1}^{m}\beta_i c^T y^i}{d_0+\sum_{i=1}^{m}\beta_i d^T y^i}=\frac{c_0+\sum_{i=1}^{m}\beta_i(-c_0)}{d_0+\sum_{i=1}^{m}\beta_i(-d_0)}=\frac{c_0}{d_0}=\text{constant}.$$

Thus we proved that if I(b) holds then $V(y)=$ constant on S_1.

CASE II: Let us denote by

$$I=\{i;\ d_i=0\},\qquad I_1=\{i;\ i\epsilon I,\ c_i>0\}.$$

Now, in Case II the following two subcases might occur

(a) $c_0+\sum_{i\epsilon I_1} c_i>0.$

(b) $c_0+\sum_{i\epsilon I_1} c_i\leqslant 0.$

If II(a) occurs $V(y)$ is not bounded from above on S_1, see [4], and if II(b) occurs (FIP) possesses an optimal solution.

An algorithm to find all the optimal solutions to (FIP) when Case II(b) or Case III (i.e. the denominator is unisignant on S_1) occur is given in section 2.

2. EXPLICIT SOLUTIONS FOR (FIP) (CASES II(b) AND III)

The optimal solutions to (FIP) when Cases II(b) or III occur are obtained in the following manner: Let

(17)
$$I=\{i;\ d_i=0\}$$

For all indices $i\epsilon I$ substitute in $V(y)$

(18)
$$y_i^*=\begin{cases}1 & \text{if } c_i>0\\ 0 & \text{if } c_i<0\\ p_i & \text{if } c_i=0 \qquad 0\leqslant p_i\leqslant 1\end{cases}$$

and denote by

(19)
$$\tilde{c}_0=c_0+\sum_{i\epsilon I} c_i y_i^*.$$

Next denote by

(20)
$$\gamma_k=\frac{c_k}{d_k}\qquad k\notin I$$

and assume for simplicity that

(21)
$$\gamma_1\geqslant\gamma_2\geqslant\ .\ .\ .\ \geqslant\gamma_r$$

where $r = m - |I|$ and $|I|$ is the number of elements in I. Let l be the smallest index for which

$$\frac{\tilde{c}_0 + \sum_{j=1}^{l-1} c_j}{d_0 + \sum_{j=1}^{l-1} d_j} \geq \gamma_l, \tag{22}$$

and let l_1 be the smallest index $l_1 > l$ for which

$$\frac{\tilde{c}_0 + \sum_{j=1}^{l_1-1} c_j}{d_0 + \sum_{j=1}^{l_1-1} d_j} > \gamma_{l^1} \tag{23}$$

In the following we shall prove that the set of optimal solutions y^{opt} to (FIP), denoted by Y, is given by

$$y^{\text{opt}} = \begin{cases} 1 & i \leq l-1 \\ p_i & l \leq i < l_1 \qquad 0 \leq P_i \leq 1 \\ 0 & l_1 \leq i \leq r \\ y_i^* & i \in I. \end{cases} \tag{24}$$

In [4], Charnes and Cooper established the optimality of a solution for (FIP), generated by (22), by exploiting the special structure of the pair of dual linear programming problems obtained after applying their well-known transformation which was introduced in [3].

An elementary proof for the optimality of Y is given in the following:

LEMMA 2: For any two fractions a_1/b_1, a_2/b_2, for which $b_1, b_2 > 0$ we have

$$\text{Min}\left[\frac{a_1}{b_1}, \frac{a_2}{b_2}\right] < \frac{a_1 + a_2}{b_1 + b_2} < \text{Max}\left[\frac{a_1}{b_1}, \frac{a_2}{b_2}\right]$$

whenever $a_1 b_2 \neq a_2 b_1$.

PROOF: Suppose $a_1/b_1 < a_2/b_2$. Then,

$$\text{Min}\ [a_1/b_1, a_2/b_2] = a_1/b_1.$$

Also,

$$\begin{aligned} a_1 \cdot b_2 < a_2 b_1 &\Rightarrow a_1 b_2 + a_1 b_1 < a_2 b_1 + a_1 b_1 \\ &\Rightarrow a_1(b_1 + b_2) < b_1(a_1 + a_2) \\ &\Rightarrow a_1/b_1 < (a_1 + a_2)/(b_1 + b_2) \\ &\Rightarrow \text{Min}\ [a_1/b_1, a_2/b_2] = a_1/b_1 < (a_1 + a_2)/(b_1 + b_2). \end{aligned}$$

Similarly, it can be shown that

$$(a_1 + a_2)/(b_1 + b_2) < \text{Max}\,[a_1/b_1,\, a_2/b_2].$$

From Lemma 2 and the manner with which (24) was constructed we conclude that the set Y consists of all the local maximum solutions to $V(y)$ over S_1. Now, since every local maximum to $V(y)$ is also a global maximum, see e.g. [6], the global optimality of the set Y follows.

REFERENCES

[1] Ben-Israel, A., and A. Charnes, "An Explicit Solution of a Special Class of Linear Programming Problems", Operations Research *16*, 1166–1175 (1968).

[2] Buhler, W., "A Note on Fractional Interval Programming", Technische Hochschule Aachen W. Germany, January 1973.

[3] Charnes, A. and W. W. Cooper, "Programming with Linear Fractional Functionals", Nav. Res. Log. Quart. *9*, 181–186 (1962).

[4] Charnes, A. and W. W. Cooper, "An Explicit Solution in Linear Fractional Programming", Nav. Res. Log. Quart. *20*, 449–467 (1973).

[5] Martos, B., "Hyperbolic Programming", translated by A. and V. Whinston, Nav. Res. Log. Quart. *11*, 135–155 (1964).

[6] Martos, B. "The Direct Power of Adjacent Vertex Programming Methods", Management Science *12*, 241–252 (1965).

A NOTE ON TESTING FOR EXPONENTIALITY*

Carl M. Harris

Department of Industrial Engineering and Operations Research
Syracuse University
Syracuse, New York

ABSTRACT

In this note some extensions are made to previous work by a number of authors on the development of tests for exponentiality. The most recent example is due to Fercho and Ringer in which they compare the small sample powers of a few well-known test statistics for the hypothesis of a constant failure rate. It is the primary intent of this current work to extend Gnedenko's F test to situations with hypercensoring and to provide guidance for its use, particularly when a log-normal distribution is the alternative.

1. INTRODUCTION

There have been numerous papers over the years on the most appropriate choice of test statistic for the hypothesis of exponentiality. A recent example is a paper by Fercho and Ringer [2], in which the authors compare the small sample powers of four well-known test statistics for the hypothesis of a constant failure rate versus the hypothesis of a nonconstant hazard. The tests were compared for samples of size $n=10(5)50$ using Weibull alternatives with shape parameters varying from 0.5 to 2.5 (thus allowing both IFRs and DFRs). The four tests used were the classical F (first posed for this problem in Gnedenko et al. [4]), two tests due to Epstein [1], and a final one due to Hartley [5]. In their conclusions the authors also mention the application of Kolmogorov-Smirnov techniques to this problem as formulated by Srinivasan [8], to which this author would like to add the work of Lilliefors [7] and a further related paper by Finkelstein and Schafer [3]. In the end, Messrs. Fercho and Ringer come to the decision that their results provide a mixed bag, though the F test seemed to have the best overall performance.

It is the primary intent of this note to extend the Gnedenko F test to situations with hypercensoring (i.e., removals occurring in a completely random fashion up until the very last failure), and to provide hints on the use of the F, particularly in the event that the alternative is a log-normal distribution.

2. THE TEST

Most methods for goodness-of-fit require complete samples (to be distinguished from those that may be censored, truncated, etc.) for their implementation. As pointed out by Fercho and Ringer, this

*This work was supported by the U.S. Office of Naval Research under Contract No. N00014-67-A-0214 when the author was with The George Washington University.

problem can be somewhat circumvented in the exponential censored case by considering the sequence of normalized spacings

$$S_i = (n-i+1)(t_i - t_{i-1}),$$

where the hypothesized exponential data points are arranged so that all n interoccurrence periods begin at time $t_0 = 0$ and t_i denotes then the time of the ith occurrence. In the reliability context, this amounts to putting all items on test at the same time and then S_i becomes the total time on test between the $(i-1)$st and ith failures. It can then be shown that under the assumption that the failure times are exponentially distributed, the normalized spacings will also be exponential, with exactly the same mean.

But this result can be made even more powerful by allowing for hypercensoring, and it turns out to be still true that the normalized spacings are indeed exponential. A proof of this contention follows (with no loss in generality) for the two-item case.

Suppose two items are put on test, each with exponential failure distribution $F(x) = 1 - e^{-\lambda x}$. Then let $Y = \min\{X_1, X_2\}$, so that

$$F_Y(y) = Pr\{\min\{X_1, X_2\} \leqslant y\} = 1 - Pr\{X_1 > y\} \cdot Pr\{X_2 > y\}$$

$$= 1 - e^{-2\lambda y}.$$

Hence Y follows the exponential distribution with parameter 2λ. Now suppose one item is censored at time x^* if it is still alive then, and let T be the *total* time on test until the first failure. So

$$T = \begin{cases} 2Y & \text{for } Y \leqslant x^* \\ x^* + X_1 & \text{for } Y > x^* \end{cases},$$

and $G_T(t) = Pr\{2Y \leqslant t \text{ and } Y \leqslant x^*\} + Pr\{x^* + X_1 \leqslant t \text{ and } Y > x^*\}$.

CASE I: $t \leqslant 2x^*$.

Here

$$Pr\{2Y \leqslant t \text{ and } Y \leqslant x^*\} = Pr\{Y \leqslant t/2\} = 1 - e^{-\lambda t}$$

and

$$Pr\{X_1 \leqslant t - x^* \text{ and } Y > x^*\} = 0.$$

Therefore

$$G_T(t) = 1 - e^{-\lambda t}.$$

CASE II: $t > 2x^*$.

Now

$$Pr\{2Y \leqslant t \text{ and } Y \leqslant x^*\} = Pr\{Y \leqslant x^*\} = 1 - e^{-2\lambda x^*},$$

$$Pr\{X_1 \leqslant t - x^* \text{ and } Y > x^*\} = Pr\{x^* < X_1 \leqslant t - x^*\} \cdot Pr\{X_2 > x^*$$

$$= e^{-\lambda x^*} \int_{x^*}^{t-x^*} \lambda e^{-\lambda s} ds,$$

$$G_T(t) = 1 - e^{-2\lambda x^*} + e^{-\lambda x^*} \int_{x^*}^{t-x^*} \lambda e^{-\lambda s} ds,$$

and

$$g_T(t) = e^{-\lambda x^*} \lambda e^{-\lambda(t-x^*)} = \lambda e^{-\lambda t}.$$

Therefore T follows the exponential distribution with parameter λ.

For this F test, the first r and last $(n-r)$ of a set of n normalized spacings (as defined earlier) from the hypothesized exponential are grouped.† Then, since the S_i are independent and identically distributed exponentials with exactly the same mean as the underlying distribution, it follows that the statistic

$$Q = \frac{\sum_{i=1}^{r} S_i/r}{\sum_{i=r+1}^{n} S_i/[n-r]}$$

is the ratio of independent Erlang variables and thus follows an F distribution with $2r$ and $2(n-r)$ degrees of freedom when the hypothesis of exponentiality is true. Therefore a two-tailed F test would be performed on the Q calculated from the set of data in order to determine whether the stream is indeed truly exponential. It should be noted that a one-tailed test is to be used when there is specific information that the alternative is IFR or DFR. This is so because an IFR (DFR) should yield a Q statistic $>(<)$ 1. Furthermore, *any* kind of information regarding the possible shape of the alternative's hazard function can generally be used to improve the power of the F test by providing a more rational means for selecting the way to divide the data in two. This is so because under the null hypothesis of exponentiality, the maximum-likelihood estimator of the hazard over any interval is simply the number of failure times falling in that interval divided by the total time on test accrued over that period of time. The usual split at $r = [n/2]$ is especially nonoptimal whenever the hazard is U-shaped, as it may be in log-normal cases, a point explored in more detail later.

3. SOME RESULTS

A number of power comparisons were made to permit the testing of the feasibility of any new approaches to the subject problem. Sample sizes were restricted to 10, 20, 30, and 40, and runs were made with 1,000 repetitions for each sample size at both the 0.05 and 0.01 levels of significance. A number of specific CDFs were selected for each alternative and results were based on a two-sided test, or a one-sided test whenever the alternative had either an IFR or DFR. A power test was also performed

†See [9] for a discussion of an interesting approach to choosing the value of r. One generally sees $r = [n/2]$ as the most reasonable choice given no additional information.

for the modified $K-S$ statistic of Finkelstein and Schafer for a log-normal alternative using the critical values contained in [3]. This variation on the usual $K-S$ theme due to Finkelstein and Schafer is called the S_n^* statistic, and is given by

$$S_n^* = \sum_{j=1}^{n} \{\max[\,|j/n - F(t_j)|, \quad |F(t_j) - (j-1)/n|\,]\}.$$

To compare these results to those of Fercho and Ringer, the reader is referred to their Table 3.3, where $\beta = 1.5$. We have, in addition to providing a run with the same value of β, also calculated the

TABLE I. *Probability of Rejecting Hypothesis of Exponentiality*

(Null hypothesis is H_0: $f(t) = e^{-t}$)

Underlying distribution	Mean	Sample size	Significance level	Estimate of power: 1-tail F	2-tail F	2-tail S_n^*
Erlang Type 2	1	10	0.05	0.216		
			0.01	0.049		
$f(t) = \lambda(\lambda t) \cdot e^{-\lambda t}$		20	0.05	0.396		
			0.01	0.121		
		30	0.05	0.573		
			0.01	0.237		
$[\lambda = 2]$		40	0.05	0.673		
			0.01	0.351		
Log-Normal	1	10	0.05	0.038		0.112
			0.01	0.004		0.017
$f(t) = \frac{1}{t\sigma\sqrt{2\pi}} \cdot \exp\left[-\frac{(\ln t - a)^2}{2\sigma^2}\right]$		20	0.05	0.143		0.174
			0.01	0.011		0.037
		30	0.05	0.032		0.201
			0.01	0.009		0.066
$\begin{bmatrix} a = -1/2 \\ \sigma^2 = 1 \end{bmatrix}$		40	0.05	0.031		0.267
			0.01	0.000		0.104
Weibull	0.903	10	0.05	0.222	0.128	
			0.01	0.070	0.040	
$f(t) = \lambda\beta t^{\beta-1} \cdot e^{-\lambda t^{\beta}}$		20	0.05	0.438	0.299	
			0.01	0.169	0.096	
		30	0.05	0.614	0.461	
			0.01	0.297	0.196	
$\begin{bmatrix} \lambda = 1 \\ \beta = 3/2 \end{bmatrix}$		40	0.05	0.773	0.635	
			0.01	0.437	0.326	

TABLE I. *Probability of Rejecting Hypothesis of Exponentiality* —Continued

(Null hypothesis is H_0: $f(t) = e^{-t}$)

Underlying distribution	Mean	Sample size	Significance level	Estimate of power: 1-tail F	2-tail F	2-tail S_n^*
Weibull	0.909	10	0.05	0.189		
			0.01	0.052		
$f(t) = \lambda\beta t^{\beta-1} \cdot e^{-\lambda t^\beta}$		20	0.05	0.368		
			0.01	0.121		
		30	0.05	0.508		
			0.01	0.214		
$\begin{bmatrix} \lambda = 1 \\ \beta = 10/7 \end{bmatrix}$		40	0.05	0.670		
			0.01	0.341		
Weibull	0.931	10	0.05	0.113	0.062	
			0.01	0.031	0.017	
$f(t) = \lambda\beta t^{\beta-1} \cdot e^{-\lambda t^\beta}$		20	0.05	0.207	0.127	
			0.01	0.042	0.026	
		30	0.05	0.275	0.160	
			0.01	0.094	0.045	
$\begin{bmatrix} \lambda = 1 \\ \beta = 5/4 \end{bmatrix}$		40	0.05	0.365	0.225	
			0.01	0.135	0.072	

simulated power when $\beta = 5/4$ and 10/7. This was done because it gives one run with a mean closer to that of the null distribution (namely, 0.931 vs. 1), a second with a mean halfway in between 1.0 and 1.5, and a third with a mean in between the other two. Furthermore, for illustrative purposes, both one and two-sided tests were used; note that Fercho and Ringer used 2,000-point (one-sided) simulations as compared to our 1,000.

It seems therefore fair to say that the current results are quite consistent with the earlier ones. We can be sure these results can all be improved by a more judicious scheme for data splitting, though the normalized spacing approach with $r = [n/2]$ tends to do pretty well.

4. SPECIALIZATION TO LOG-NORMAL

The log-normal alternative turns out to be a very special case, since the $r = [n/2]$ rule for data splitting is especially unwise. A careful look at the plot of a log-normal with mean 1, for example,

$$f(t) = \frac{1}{t\sqrt{2\pi}} \exp - \left[\frac{(\ln t + 1/2)^2}{2} \right]$$

and its hazard rate

$$h(t) = \frac{f(t)}{1-F(t)}$$

$$= \frac{f(t)}{\frac{1}{\sqrt{2\pi}} \int_{\ln t + 1/2}^{\infty} e^{-x^2/2}\, dx},$$

gives extremely important information which can be used to greatly increase the power of the F test. We have chosen to split the data in such a way that the numerator of the statistic is based on the first $[n/4]$ points plus the last $[n/4]$, while the denominator is calculated from the remaining points. This simple rule increases the power when the sample size is 30 at $\alpha = 0.05$ from 0.032 to 0.553, a very significant improvement, which now allows the F to be superior to S_n^*. The usual $([n/2], n-[n/2])$ split does not work well in this case because the shape of the hazard tends to be decreasing for the first half of the data and increasing for the second half with the two averages working out to be very close to each other. Any such direct use of information regarding the shape of the alternative's hazard rate can always be used as an easy but important way of improving the F's power.

TABLE II. *Comparison of Results*

n	Level	Fercho & Ringer	Harris: $\beta = 1.5$		Harris: $\beta = 1.25$	
			1-tail	2-tail	1-tail	2-tail
10	0.05	0.126	0.222	0.128	0.113	0.062
	0.01	0.028	0.070	0.040	0.031	0.017
20	0.05	0.302	0.438	0.299	0.207	0.127
	0.01	0.112	0.169	0.096	0.042	0.026
30	0.05	0.455	0.614	0.461	0.275	0.160
	0.01	0.206	0.297	0.196	0.094	0.045
40	0.05	0.620	0.773	0.635	0.365	0.225
	0.01	0.320	0.437	0.326	0.135	0.072

TABLE III. *Results for Log-Normal With Modified Spacing Plan*

n	Level	1-tail F	Previous 1-tail F	S_n^*
30	0.05	0.553	0.032	0.201
	0.01	0.149	0.009	0.066

5. CONCLUSIONS

In summary then, though some other test statistics have some competitive advantages when

testing against specific alternatives, the F test comes out with the best net performance. Beyond all the pluses already mentioned, such as the handling of censoring, there is the further advantage of ease of computation and ready availability of critical points. In fact, it is rather easy to store F critical values for all possible pairs of degrees of freedom using one of many numerical approximation procedures from the literature (for example, see [6]).

REFERENCES

[1] Epstein, B. "Testing for the validity of the assumption that the underlying distribution of life is exponential." Technometrics *2*, 83–101 (1960).

[2] Fercho, W. W. and L. J. Ringer, "Small sample power of some tests of constant failure rate." Technometrics *14*, 713–724 (1972).

[3] Finkelstein, J. and R. E. Schafer, "Improved goodness-of-fit tests." Biometrika *58*, 641–645 (1971).

[4] Gnedenko, B. V., Y. K. Belyayev, and A. D. Solovyev, *Mathematical Methods of Reliability Theory.* (Academic Press, New York, 1969).

[5] Hartley, H. O. "The maximum F ratio as a short-cut test for homogeneity of variance." Biometrika *37*, 308–312 (1950).

[6] Johnson, E. E. "Empirical equations for approximating tabular F values." Technometrics *15*, 379–384 (1973).

[7] Lilliefors, H. W. "On the Kolmogorov-Smirnov test for the exponential distribution with mean unknown." J. Amer. Statist. Assoc. *64*, 387–389 (1969).

[8] Srinivasan, R. "An approach to testing the goodness of fit of incompletely specified distributions." Biometrika *57*, 605–611 (1970).

[9] Stollmack, S. and C. M. Harris "Failure-rate analysis applied to recidivism data." Operations Res. *22*, 1192–1205 (1974).

NEWS AND MEMORANDA

THE 1975 LANCHESTER PRIZE
Call for Nominations

Each year since 1954 the Council of the Operations Research Society of America has offered the Lanchester Prize for the best English-language published contribution in operations research. The Prize for 1975 consists of $2,000 and a commemorative medallion.

The screening of books and papers for the 1975 Prize will be carried out by a Committee appointed by the Council of the Society. To be eligible for consideration, the book or paper must be nominated to the Committee. Nominations may be made by anyone; this notice constitutes a call for nominations.

To be eligible for the Lanchester Prize, a book, a paper, or group of papers must meet the following requirements:

(1) it must be on an operations research subject,

(2) it must carry a 1975 publication date,

(3) it must be written in the English language, and

(4) it must have appeared in the open literature.

The book or paper(s) may be a case history, a report of research representing new results, or primarily expository.

Judgments will be made by the Committee using the following criteria:

(1) the magnitude of the contribution to the advancement of the state of the art of operations research,

(2) the originality of the ideas or methods,

(3) new vistas of application opened up,

(4) the degree to which unification or simplification of existing theory or method is achieved, and

(5) expository clarity and excellence.

Nominations should be sent to:

Richard C. Larson, Chairman
1975 Lanchester Prize Committee
Room 4–209
Massachusetts Institute of Technology
Cambridge, MA 02139.

Nominations may be in any form, but must include as a minimum the title(s) of paper(s) or book, author(s), place and date of publication. Supporting statements bearing on the worthwhileness of the publication in terms of the five criteria will be helpful, but are not required. Each nomination will be carefully screened by the Committee, but nominations should be received by June 1, 1976 to allow time for adequate review.

Announcement of the results of the Committee and ORSA Council action, as well as award of any prize(s) approved, will be made at the 50th National Meeting of the Society, November 3–6, 1976, in Miami Beach, Florida.

PRICE INCREASE FOR SUBSCRIBERS

The cost of a one year subscription to the Naval Research Logistics Quarterly is now $11.15 a year in the U.S. and Canada, and $13.95 elsewhere. Subscriptions may be ordered by writing to the Superintendent of Documents, U.S. Government Printing Office, Washington, D.C. 20402.

NAVAL RESEARCH LOGISTICS QUARTERLY

The Naval Research Logistics Quarterly is devoted to the dissemination of scientific information in logistics and will publish research and expository papers, including those in certain areas of mathematics, statistics, and economics, relevant to the over-all effort to improve the efficiency and effectiveness of logistics operations.

Information for Contributors is indicated on inside back cover.

The Naval Research Logistics Quarterly is published by the Office of Naval Research in the months of March, June, September, and December and can be purchased from the Superintendent of Documents, U.S. Government Printing Office, Washington, D.C. 20402. Subscription Price: $11.15 a year in the U.S. and Canada, $13.95 elsewhere. Cost of individual issues may be obtained from the Superintendent of Documents.

NAVAL RESEARCH LOGISTICS QUARTERLY

MARCH 1976
VOL. 23, NO. 1
NAVSO P-1278

CONTENTS

ARTICLES

OFFICE OF NAVAL RESEARCH
Arlington, Va. 22217